The Dissertation Committee for Chi-san Ho certifies that this is the approved version of the following dissertation:

Mixtures of Triangular Densities with Applications to Bayesian Mode Regressions

Committee:

Paul Damien, Supervisor

Carlos Carvalho

Prabhudev Konana

Tom Sager

Tom Shively

Stephen Walker

Mixtures of Triangular Densities with Applications to Bayesian Mode Regressions

by

Chi-san Ho, B.B.A.; M.S.Stat; M.S.I.R.O.M.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2014

Acknowledgements

First of all, I would like to thank my advisor, Prof. Paul Damien, for his support in both intellectual and emotional ways. He made my entire PhD experience much better and I am really grateful and feel extremely lucky to be his student. I would also like to thank my committee members: Prof. Carvalho, Prof. Konana, Prof. Sager, Prof. Shively and Prof. Walker for their valuable input and suggestion regarding to my thesis work. Many thanks to Prof. Gu and my fellow graduate student Alvin Leung for their support in the stock sentiment project.

Next, I would like to thank Prof. Betsy Greenberg. I started to be her teaching assistant the first year I attended UT. It was a pleasure to work for her and she also gave me her prospect as a female professor when I encounter career choice.

Much thanks to Ms. Hillary Patterson and Ms. Caroline Walls for their administrative help. Especially thanks to Caroline for assisting me with all the paper work and deadlines. Her resourceful and prompt support makes the entire defense process smooth.

My next appreciation goes to Austin Chinese Choir and our former conductor Ms. Chung-Hwa Chen. During my stay in Austin, all choir members took care of me as one of their own family. We have practiced together from the 10th year anniversary concert to 15th year anniversary concert. My life as a PhD student is unbreakable with my participation as a choir member.

I would like to thank all my fellow graduate students whom I had the opportunity to share my live and research experiences with. Special thank to Vivek Vasudeva for all the extraneous discussions regarding different topics except for our own research!

I would like to thank You-Know-Who for everything she did for me. Hope this will be a nice ending even through we have caused so much trauma to each other. I know you will read this and I want you to know that I do appreciate you.

Finally I would like to thank my mom for her unconditional support.

Mixtures of Triangular Densities with Applications to Bayesian Mode Regressions

Chi-san Ho, Ph.D.

The University of Texas at Austin, 2014

Supervisor: Paul Damien

Abstract

The main focus of this thesis is to develop full parametric and semiparametric Bayesian inference for data arising from triangular distributions. A natural consequence of working with such distributions is it allows one to consider regression models where the response variable is now the mode of the data distribution. A new family of nonparametric prior distributions is developed for a certain class of convex densities of particular relevance to mode regressions.

Triangular distributions arise in several contexts such as geosciences, econometrics, finance, health care management, sociology, reliability engineering, decision and risk analysis, etc. In many fields, experts, typically, have a reasonable idea about the range and most likely values that define a data distribution. Eliciting these quantities is thus, generally, easier than eliciting moments of other commonly known distributions. Using simulated and actual data, applications of triangular distributions, with and without mode regressions, in some of the aforementioned areas are tackled.

Contents

List of Figures

List of Tables

1 Introduction

1.1 The Triangular Distribution

The triangular distribution is generally used to describe a population for which there is limited sample data, but for which reasonable guesses at its minimum, maximum and mode are available. Since practitioners, typically, have such knowledge, the triangular distribution is used in project management, including popular tools such as Program Evaluation Review Technique (PERT) and Critical Path Method (CPM).

Triangular distributions are most widely used in oil and gas exploration where data are expensive to collect, and where it is difficult to accurately model the population being sampled. Zhang (2003) describes the role of these distributions in geosciences where the use of subjective prior knowledge plays a prominent role when compared to data rich contexts. Floris et al. (2001) and Barker et al. (2001) develop pseudo Bayesian models in the petroleum industry using importance sampling methods. This modeling process is a crude way of using Bayesian ideas to validate existing data; the latter are extremely expensive to collect. Triangular distributions are used to construct the prior models.

Rao (2010) uses the triangular distribution in the following example to illustrate its use in risk analysis. The goal is to determine the volume of oil one could recover from an underground reservoir. Executing 3D seismic surveys could cost upwards of \$50,000 a day and exploratory wells could cost between \$1 million and \$20 million. Any estimate of reserves is obtained using a limited number of data points; indeed, to validate their Gaussian pseudo-Bayesian simulation models, Floris et al. (2001) work with *one* observed data point! Hence any uncertainty quantification using these models is likely to be very poor, since a small number of estimates obtained from exploratory wells may not be representative of the entire oil field. Continuing with the Rao (2010) example, one of the key parameters in simulation models in petroleum engineering is porosity of rock formations. Suppose an experimentally observed value for this is 10%. But a geologist might treat this as a realization from a triangular distribution with minimum, modal and maximum values of, say 2.5%, 7.5% and 20%. A similar process could be applied to other critical parameters, namely the area of the field, its thickness, and recovery factor. The total volume in the field can then be calculated using:

$$V = A \times T \times P \times R,$$

where V is the volume of oil (in metric tonnes) to be recovered from the reservoir; A is the area of the reservoir in squared kilometers KM^2; T is the thickness of the reservoir in meters M; P is the porosity, which is the age (in percent) of the reservoir's rock volume that is void space. For example, if a reservoir has a porosity of 10%, the void space in 1 cubic meter of rock which might contain fluid is 0.1 cubic meter. R is the recovery factor. Since it is unlikely that all the oil from the reservoir will be recovered, R provides an estimate of the age (in percent) that could be extracted from the reservoir. Given considerable uncertainty in the four parameters, subjective triangular distributions are assigned. Monte Carlo simulations from these distributions are executed thousands of times to produce a distribution for the total volume, V, in the field; this volume distribution is typically unimodal, asymmetric and could also be described via a triangular distribution whose parameters are estimated from the Monte Carlo samples.

Patel et al. (2011) use the triangular distribution to construct a genetic algorithm for resource estimation, namely a process by which the economically recoverable hydrocarbons within a reservoir are calculated.

The statistical literature on triangular densities is sparse. Law and Kelton (2000) is the only book where the triangular distribution is discussed in considerable detail, largely due to its significant role in risk management. Scherer et al. (2003) develop theory and computation to show that the normal and log-normal distributions can be closely and easily approximated using triangular densities.

Perron and Mengersen (2001) first applied mixtures of triangular distributions in a nonparametric Bayesian context. Their work considers nonparametric estimation of a monotone increasing function and its use in survival analysis. This is similar in spirit to the Bayesian models in Smith and Kohn (1997) and Wood and Kohn (1998). McVinish et al. (2009) also research mixtures of triangular distributions. They consider consistency of Bayes factors in goodness of fit testing; to accomplish this task, they adapt the results from Perron and Mengersen (2001). One point to be noted in these applications of triangular mixtures is that the components in these mixtures of triangular distributions are located at different modes and, therefore, the mixtures are not necessarily unimodal.

In this thesis, a special form of triangular densities using mixtures of Dirichlet processes (MDP) is used, where all components share the same mode at 0. As a result, mixtures of triangular densities can be used to approximate a wide class of unimodal, symmetric or asymmetric, densities with different ranges for kurtosis. Specifically, this thesis will show that the class of mixtures of triangular densities (MTD) covers those unimodal densities of specific relevance to mode regressions. Mixtures of normals

are not guaranteed to be unimodal and mixtures of uniforms do not cover unimodal and asymmetric densities at the same time.

1.2 Mode Regressions

It is well-known that the arithmetic mean may not be appropriate as a measure of central tendency if the data are skewed or if they contain outliers. It is also true that the mean and median of two densities may be identical while the shapes may be different. Mode preserves key features in the underlying density of a population (such as wiggles) when the mean and median may "smooth" out the data. It is for this reason, mode has been used in network systems; see Hedges and Shah (2003), Heckman et al. (2001), Kumar and Hedges (1998), Markov et al. (1997).

Mode estimation using nonparametric kernel methods has been studied by various authors, including Yasukawa (1926), Parzen (1962), Chernoff (1964), Grenander (1965), Eddy (1980), Bickel and Fan (1996), Birg'e (1997), Berlinet et al. (1998) and Meyer (2001).

Another strand of literature involves conditional mode estimation using nonparametric conditional density estimation; see, Collomb et al. (1987), Samanta and Thavaneswarn (1990), Quintela-Del-Rio and Vieu (1997), Ziegler (2003), Gasser et al. (1998), Hall and Huang (2001), Hall et al. (2001), Brunner (1992), Ho (2006), Dunson et al. (2007). But these papers do not provide a direct estimate of the conditional mode.

In the econometrics literature, direct inference for mode regression was first tackled by Lee (1989,1993). Using density estimation techniques (Silverman 1986), Lee (1989) developed a rectangular as well as a quadratic mode regression model under a well-known loss function that uses a rectangular or uniform kernel. In this model, the expectation is minimized at mode$(y|x)$ under the specification, mode$(y|x) = x'\beta$. Kemp and Silva (2012) proposed a semi-parametric mode regression estimator for the case in which the dependent variable has a continuous conditional density with a well-defined global mode. Unfortunately, these mode regression estimators are of little practical use since they are generally intractable; see Kim and Pollard (1990) and Kemp and Silva (2012). One key point from Lee's papers is that for the slope coefficients to be asymptotically consistent, the conditional distribution of the data, given the regressors, has to be unimodal and symmetric about the mode.

Kemp and Silva (2012) note that mode regression is useful in many applications such as wage distributions, pricing theory, energy intake, etc, where the mode is generally located below the mean and median. In other words, mean and median regressions would convey very little information about

the mode in such instances. Likewise, quantile regressions generally fail to reveal any information about the conditional mode of the data distribution. Kemp and Silva construct examples where the mean and all the quantiles are increasing functions of a regressor, but the mode decreases with the same regressor. They proceed to develop a mode regression model for fully observed, unbounded, continuous random variates with a strict unimodal conditional density. Like Lee (1989), their approach is also semiparametric and relies on the use of smooth, unbounded kernels. In a Working Paper version of their paper, Kemp and Silva demonstrate the value of their mode regressions using simulated data, and they also consider a real data sociological study. Specifically, they research the recent evolution of the body-mass index (BMI) in England, using survey data from the Health Survey of England. They model the conditional distribution of BMI as a function of year, gender, race and age. One key finding from their analysis is that the mode of the conditional distribution of BMI for females decreases over time, which is in sharp contrast to the mean and median regressions for the same data. They conclude mode regressions can provide key information about how regressors influence the location and shape of the conditional distribution of BMI, unlike traditional mean and quantile regressions. But, like Lee (1989), Kemp and Silva's approach to mode regression also suffers from implementation issues; for instance, coefficient estimation has poor convergence rates, bandwidth selection is somewhat arbitrary, and they only provide approximate normal confidence intervals.

In many e-commerce and finance applications involving portfolio allocations, the data are generally unimodal, asymmetric, and with high kurtosis. For example, Hong et al. (2007) found strong evidence of asymmetries for both size and momentum portfolios. Thus, Bayesian mode regressions could be useful in estimating and evaluating these portfolios.

All the papers cited above use classical statistical methods. The Bayesian literature on mode regressions, like the literature on mixtures of triangular distributions (MTD), is somewhat limited. Perron and Mengersen (2001) use triangular mixtures to develop nonparametric Bayesian estimates for increasing functions and illustrate their approach for hazard rate models in survival analysis. Their approach is very different than the one developed in this paper, and does not cover mode regressions with asymmetric errors. Additionally, we construct a new family of nonparametric prior distributions using MTD.

An interesting and useful application of triangular mixtures is described in Cai et al. (2008). These authors use these mixtures to devise an efficient, adaptive Metropolis-Hastings algorithm. A recent paper by Yu and Aristodemou (2014) introduces a Bayesian framework for direct mode regression in-

ference using three approaches, namely a parametric, nonparametric and an empirical likelihood based model. For the parametric Bayesian model, they use a likelihood function based on a mode uniform distribution. They prove that posterior estimates of the regression parameters based on this likelihood, even under misspecification, are consistent and asymptotically normal. For the nonparametric Bayesian model they use Dirichlet process mixtures of mode uniform distributions. Another result in their work is that for a variety of improper priors for the unknown model parameters, a proper posterior joint distribution can be derived. However, the mixture of symmetric (or mode) uniform densities cannot approximate asymmetric densities and, as a result, their mode regressions are also mean regressions.

The proposed scope of this thesis is now listed.

- This thesis is the first attempt at developing theory for nonparametric Mixtures of Triangular Densities (MTD) using a stick-breaking version of the Dirichlet Process prior; thus, a very wide family of unimodal, symmetric or asymmetric, densities with varying degrees of kurtosis can be modeled. Specifically, since mode regressions serve as the main application for this research, a new family of nonparametric prior distributions is developed to index a class of convex densities of particular relevance to mode regressions. Key theorems punctuating this aspect of the research will be proved.

- The new family of MTD will be applied to Bayesian mode regressions where novel Markov chain Monte Carlo methods will be developed to sample the posterior and predictive distributions of interest.

- There is a rich literature on dynamic Bayesian mean regression models; see, for example, West and Harrison (1997) and Carlin et al. (1992). This thesis proposes to develop dynamic mode regressions using state space representations wherein MTD are used to model the stochastic error in the observation equation. Computational algorithms to implement this new class of dynamic mode regression models will be developed.

- The new models and methods will be exemplified using simulated and real data.

2 The Triangular Distribution and MTD

The aim of this chapter is to develop relevant theory needed to implement Dirichlet Process Mixtures of Triangular Distributions (MTD). To this end, triangular distributions need to discussed.

2.1 The Triangular Density

The triangular distribution is a continuous probability distribution with lower limit b_1, upper limit b_2 and mode c, where $b_1 \leq c \leq b_2$. The probability density function is given by

$$f(x|b_1,\, b_2,\, c) = \begin{cases} 0 & \text{for } x < b_1 \\ \frac{2(x-b_1)}{(b_2-b_1)(c-b_1)} & \text{for } b_1 \leq x < c \\ \frac{2(b_2-x)}{(b_2-b_1)(c-b_1)} & \text{for } c \leq x \leq b_2 \\ 0 & \text{for } x > b_2 \end{cases}$$

However, this form is not compatible for developing Bayesian models. In the following, a new paramaterization of the triangular density is described to facilitate Bayesian inference. To motivate this parametrization, it is instructive to first consider a special type of Gibbs sampler.

2.1.1 Constructing Densities via Auxiliary Variables

A Gibbs sampler requires that the conditional distributions for each parameter of interest is known in advance and that one can sample from these conditionals in a reasonable manner. Trouble occurs in a Gibbs sampler when a conditional distribution is not a well-known distribution and cannot be easily sampled. In such instances, one must turn to alternative computationally intensive sampling methods, like rejection sampling. But it is well-known that methods like rejection, Metropolis-Hastings, etc. have issues such as poor acceptance rates, tuning, choosing dominating densities, and so forth.

Following Besag and Green (1993), Higdon (1998), Damien et al. (1999) and Neal (2003), the main idea is to construct joint densities by introducing auxiliary variables with the aim of developing marginal distributions easily. The auxiliary, or latent, variables must be introduced in such a way that their presence properly preserves the original density function. The real advantage of this so-called "slicing method" is that after introducing the latent variables, the conditional distribution of each parameter is often simple to sample, avoiding computationally infeasible sampling methods. Damien

et al. (1999) prove that given a target density $f(x)$, one can construct a joint density $f(x,u)$, where u is a latent variable. The choice of $f(u)$ is such that upon integrating u, one recovers $f(x)$. Choosing $f(u)$ is arbitrary and requires a bit of trial and error in some instances; most of the time, however, the choice is apparent. The key is that the full conditional distributions in the resulting Gibbs sampler are, typically, of known type.

2.1.2 An Auxiliary Variable Construction of a Triangular Distribution

Consider the following:

THEOREM 1: Let v be a random variable with parameter $a > 0$; assume its density function is given by:

$$f(v) = \frac{2v}{a^2}, \text{ where } 0 < v < a.$$

Let the conditional distribution of the data, say y, given v, follow a uniform distribution with parameter $\lambda \geq 0$:

$$f(y|v) = uniform(-\exp(-\lambda) \cdot v, \exp(\lambda) \cdot v) = \frac{1}{2\cosh(\lambda) \cdot v}.$$

Then, the marginal distribution $f(y)$ is triangular with mode at zero.

PROOF: Consider the joint distribution of (y,v):

$$f(y,v) = f(v) \cdot f(y|v) = \frac{v/a^2}{v \cdot \cosh(\lambda)} \mathbf{1}(0 < v < a) \cdot \mathbf{1}(-\exp(-\lambda) \cdot v < y < \exp(\lambda) \cdot v). \tag{2.1}$$

Intergrate out v,

$$\int f(y,v)dv = \int \frac{\mathbf{1}\left(-e^{-\lambda}v < y < e^{\lambda}v\right)\mathbf{1}(0 < v < a)}{\cosh(\lambda)\, a^2} dv$$

$$= \int_0^a \frac{\mathbf{1}\left(-e^{-\lambda}v < y < e^{\lambda}v\right)}{\cosh(\lambda) \cdot a^2} dv = \int_{\max(ye^{-\lambda}, -ye^{\lambda})}^a \frac{1}{\cosh(\lambda) \cdot a^2} dv$$

$$= \frac{a - \max\left\{ye^{-\lambda}, -ye^{\lambda}\right\}}{\cosh(\lambda) \cdot a^2}.$$

As a result, we have

$$f(y) = \frac{a - \max\{y \cdot \exp(-\lambda), -y \cdot \exp(\lambda)\}}{\cosh(\lambda) \cdot a^2}, \quad -a \cdot \exp(-\lambda) < y < a \cdot \exp(\lambda) \tag{2.2}$$

Now, any triangular density with mode 0, lower limit $b_1 < 0$ upper limit $b_2 > 0$ can be reparameterized to coincide with $f(y)$ with parameters $(a,\ \lambda)$. And so, let $-a \cdot \exp(-\lambda) = b_1$ and $a \cdot \exp(\lambda) = b_2$, then $a = \sqrt{-b_1 b_2}$ and $\lambda = \frac{\ln(b_2) - \ln(-b_1)}{2}$; after some algebra, we obtain:

$$f(y|a,\ \lambda) = \begin{cases} 0 & \text{for } y < -a \cdot \exp(-\lambda) = b_1 \\ \frac{a + y \cdot \exp(\lambda)}{\cosh(\lambda) \cdot a^2} = \frac{2(y - b_1)}{-b_1(b_2 - b_1)} & \text{for } b_1 = -a \cdot \exp(-\lambda) < y < 0 \\ \frac{a - y \cdot \exp(-\lambda)}{\cosh(\lambda) \cdot a^2} = \frac{2(b_2 - y)}{-b_1(b_2 - b_1)} & \text{for } 0 < y < a \cdot \exp(\lambda) = b_2 \\ 0 & \text{for } y > b_2 \end{cases} \quad \square$$

In the above theorem, the value of λ indicates the level of skewness; if λ deviates from 0, the distribution of y becomes more asymmetric. When $\lambda = 0$, the distribution of y is symmetric. The length of the tail of the distribution of y is determined by the size of a; so this acts as the kurtosis parameter.

2.1.3 Gibbs Sampler for the Triangular Distribution

Theorem 1 is also useful from a computational perspective since it leads to a Gibbs sampler (Gelfand and Smith 1990) that is easy to implement. Hence, for a single observation y from $f(y)$, consider the following Gibbs sampler, where one has to sample from the conditional distributions of a, λ and the auxiliary variable v. Throughout, we work with equation (2.1). Also, where appropriate, $\pi(.)$ denotes a prior distribution; we have:

$$f(v|y, a, \lambda) \propto f(v|a) \cdot f(y|v, \lambda) = \frac{1}{a^2 \cdot \cosh(\lambda)} \mathbf{1}\left\{\max\left(y \cdot \exp(-\lambda), -y \cdot \exp(\lambda)\right) < v < a\right\}.$$

$$f(v|y, a, \lambda) = \mathbf{1}\{\max(y \cdot \exp(-\lambda), -y \cdot \exp(\lambda)) < v < a\} \tag{2.3}$$

From equation (2.1), it is clear that the posterior conditional distribution of a only depends on v:

$$f(a|v, y, \lambda) \propto \pi(a) \cdot f(v|a) \propto \pi(a) \cdot \frac{v}{a^2} \mathbf{1}(a > v).$$

Take $\pi(a)$ to be Pareto(α_1, α_2), then the posterior conditional distribution of a is Pareto($\max(\alpha_1, v)$, $\alpha_2 + 2$).

For the conditional distribution of λ, from equation (2.1),

$$f(\lambda|y, v, a) \propto \pi(\lambda) \cdot f(y|v, \lambda) \propto \pi(\lambda) \cdot \frac{1}{\cosh(\lambda)} \mathbf{1}\left(\ln\left(\frac{y^+}{v}\right) < \lambda < \ln\left(\frac{-v}{y^-}\right)\right)$$

where y^+ and y^- indicates $y > 0$ and $y < 0$, respectively. Unlike a, there is no convenient prior choice for λ to obtain a recognizable density. Thus, to sample from the above conditional, one could try different methods. For this simple model, one could use a Metropolis-Hastings algorithm with a Gaussian proposal with mean 0.

Now suppose we observe $Y = (y_1, \cdots, y_n)$ from $f(y)$. To start the Gibbs sampler, careful choices of initial values for a and λ are needed since these parameters are all bounded. Finding feasible initial values could be tricky; this point is further discussed when the model is illustrated with simulated data.

Given initial values for a and λ, for each y_i we can sample from the conditional distribution of v_i according to (2.3); denote these samples as $\mathbf{v} = \{v_1, ..., v_n\}$, with which we obtain:

$$f(a|\mathbf{v}, Y, \lambda) \propto \pi(a) \cdot \prod_i f(v_i|a) \propto \pi(a) \cdot \frac{1}{a^{2n}} \mathbf{1}(a > \max v_i). \tag{2.4}$$

With a Pareto prior, the posterior conditional distribution of a is Pareto($\max(\alpha_1, v_m)$, $\alpha_2 + 2n$), where $v_m = \max(\mathbf{v})$.

Finally, sampling from the conditional distribution of λ, given Y and $\mathbf{v}$ involves generating samples from the following:

$$f(\lambda|Y, \mathbf{v}, a) \propto \pi(\lambda) \cdot \prod_i f(y_i|v_i, \lambda)$$

$$\propto \pi(\lambda) \cdot \left(\frac{1}{\cosh(\lambda)}\right)^n \mathbf{1}\left(\max\left(\ln\left(\frac{y_i^+}{v_i}\right)\right) < \lambda < \min\left(\ln\left(\frac{-v_i}{y_i^-}\right)\right)\right) \tag{2.5}$$

As noted earlier, sampling this last conditional distribution is possible using, say a Metropolis-Hastings algorithm.

2.2 The Mixture of Triangular Densities

In this section, we develop the theory and computation for the mixture of triangular densities (MTD).

2.2.1 A Nonparametric Auxiliary Variable Construction of MTD

Lo (1984) introduced the now famous Mixture of Dirichlet Process (MDP) models; specifically, with a Gaussian kernel, the MDP model is given by

$$f_P(y) = \int N\left(y; \mu, \sigma^2\right) dP(\phi),$$

where $P \sim D(M, P_0)$ is Ferguson's (1973) Dirichlet Process with scale parameter $M > 0$ and P_0 is a baseline distribution with $\phi = (\mu, \sigma^2)$, namely the mean and variance of the normal distribution. There is a plethora of papers that use the MDP model since modern MCMC methods can be used to do full Bayesian inference, after noting the fact that it is possible to integrate out P from the posterior distribution obtained from this model; see Kalli et al. (2011).

Following Sethuraman (1994), it is possible to represent $P \sim D(M, P_0)$ using a stick-breaking representation:

$$P = \sum_{j=1}^{\infty} \omega_j \kappa_{\phi_j}$$

where ϕ_1, ϕ_2, $\phi_3 \cdots$ are independent and identically distributed (i.i.d.) from P_0 and

$$\omega_1 = \delta_1, \quad \omega_j = \delta_j \prod_{l<j} (1 - \omega_l)$$

where the δ_js are i.i.d from a Beta distribution with parameters $(1, M)$, denoted $\text{Be}(1, M)$.

Kalli et al. (2011) argue that the Sethuraman representation is critical in estimating mixture models; see, also, Walker (2007). Here this insight is used to write the MTD representation as follows:

$$f(y) = \sum_j w_j \frac{a_j - \max\left\{y e^{-\lambda_j}, -y e^{\lambda_j}\right\}}{2\cosh(\lambda_j) \cdot a_j^2}, \quad \min\left(-a_j e^{-\lambda_j}\right) < y < \max\left(a_j e^{\lambda_j}\right). \tag{2.6}$$

In the next subsection, the above equation will be shown to be a prior on a space of specific convex densities of particular relevance to mode regressions.

2.2.2 MTD and Convex Densities

Theorem 2 below shows that MTD with mode zero are convex on both sides of the mode. This, in turn, is used to motivate Theorem 3 and the subsequent Corollary whose upshot is that MTD with mode zero can approximate any unimodal, untruncated, convex density, thus serving as a nonparametric prior on this family of densities. In Chapter 4 these results will be used to model mode regressions.

Definition: Unimodal, Untruncated Convex (UUC) Densities:

Continuous nimodal densities convex on both sides of the mode are untruncated when the bounds exist; at the bounds, the densities are zero.

THEOREM 2: Suppose y is distributed MTD with modes at zero. Then the density function of y is convex on the interval $y > 0$ and on the interval $y < 0$. Conversely, any piecewise linear density function $h(x)$ with mode at 0 and convex on each side of the mode can be represented by an MTD.

PROOF: For $y > 0$, we can express the mixture density as

$$f(y) = \sum_j w_j \cdot \max\left(a_j - b_j y,\, 0\right)$$

where $a_j > 0$ and $b_j > 0$; this is equivalent to the MTD representation in equation 2.6 obtained via Theorem 1.

For each component, the density fuction is convex on $y > 0$. Since the sum of two convex functions is also convex and this property can be extended to infinite sums, the MTD is convex on the interval $y > 0$. A similar argument applies to the interval $y < 0$.

To show that any piecewise linear function $h(x)$ with mode at 0 and convex on each side of the mode can be represented by an MTD, first consider $x > 0$; since $h(x)$ is a density function,

$$\int_0^\infty h(x)dx = q \leq 1.$$

Let $\{(x_j,\, y_j),\, -\infty < j < \infty\}$ be the vertices of $h(x)$ and let $m_j = \frac{(y_{j+1} - y_j)}{(x_{j+1} - x_j)}$ be the corresponding slope of $h(x)$ for $x_j < x < x_{j+1}$.

Since $h(x)$ is convex, $m_{j-1} < m_j < m_{j+1} < ... < 0$. Define $d_j = m_j - m_{j+1}$ and $m_j = \sum_{i>j} d_i$ then

$$h(x) = y_j - m_j x_j + m_j x,\; x_j < x < x_j + 1$$

$$\int_0^\infty h(x)dx = \sum_j \frac{(x_{j+1} - x_j)(y_{j+1} + y_j)}{2}.$$

Note that $y_j = \sum_{i=j}^\infty (y_i - y_{i+1}) = \sum_{i=j}^\infty m_i (x_i - x_{i+1}) = \sum_{i=j}^\infty \left(\sum_{k>i} d_i\right)(x_i - x_{i+1}) = \sum_{k>i} d_i (x_i - x_k)$.

Define $MT_1(x) = \sum w_i \cdot g_i(x)$ for $x > 0$ and $\sum w_i = q$ where

$$g_i(x) = \max\left(\frac{2}{x_i} - \frac{2}{x_i^2}x,\, 0\right) \text{ and } w_i = \frac{-d_i x_i^2}{2}.$$

$$MT_1(x) = \sum w_i \cdot g_i(x) = \sum_i \frac{-d_i x_i^2}{2}\left(\frac{2}{x_i} - \frac{2}{x_i^2}x\right)\mathbf{1}(0 < x < x_i)$$

$$= \sum_i (-d_i x_i + d_i x,\, 0)\,\mathbf{1}(0 < x < x_i).$$

Then for $x_j < x < x_{j+1}$,

$$MT_1(x) = \left(\left(\sum_{i>j} -d_i x_i\right) + \sum_{i>j} d_i x\right) = \left(\left(\sum_{i:\, x_i > x_j} -d_i x_i\right) + m_j x\right)$$

$$= \left(\left(\sum_{i>j} -d_i x_j - d_i(x_i - x_j)\right) + m_j x\right) = \left(\left(\sum_{i>j} -d_i(x_i - x_j)\right) - m_j x_j + m_j x\right) = y_j - m_j x_j + m_j x$$

Thus

$$MT_1(x) = h(x) \text{ for } x > 0.$$

Since

$$\int_0^\infty g_i(x)dx = 1,$$

$$\int_0^\infty h(x)dx = \int_0^\infty MT_1(x)dx = \sum w_i \int_0^\infty g_i(x)dx = \sum w_i = q.$$

Similarly, for $x < 0$, $\exists\ k_i(x)$ and $\delta_i > 0$ s.t.

$$MT_2(x) = \sum \delta_i \cdot k_i(x) \text{ for } x < 0 \text{ and } \sum \delta_i = 1 - q.$$

Combining both sides gives

$$h(x) = MT_1(x) + MT_2(x) \text{ and } \sum w_i + \sum \delta_i = 1 \qquad \square$$

COROLLARY: Let C be the collection of UUC densities. Then any $f \in C$ can be represented as an infinite MTD.

PROOF: Since any convex density on the real line can be approximated by piecewise linear functions (see pages 13-20 of Bannerman-Thompson (2008)), the result follows from theorems 2 and 3 above.

Thus, the class of MTD serves as a nonparametric prior distribution on the space of UUC densities.

THEOREM 3: The MTD are a conjugate prior on the space of UUCs.

PROOF: This follows from the stick-breaking construction of the Dirichlet process in equation (2.6).

2.2.3 Gibbs Sampler for MTD

Here we develop the Gibbs Sampler for MTD defined in equation (2.6). The goal is to determine which finite number of variables in the mixture need sampling, resulting in a valid Markov chain with the right stationary distribution. To achieve this goal, we once again employ an auxiliary variable v to equation (2.6):

$$f(y,v) = \sum_j 2w_j \frac{\mathbf{1}\left(-e^{-\lambda_j} v < y < e^{\lambda_j} v\right) \cdot \mathbf{1}\left(0 < v < a_j\right)}{2\cosh\left(\lambda_j\right) \cdot a_j^2}. \tag{2.7}$$

CLAIM: Integrating out v returns the density of y, $f(y)$. To validate this, consider the following steps.

$$f(y) = \int f(y,v)dv = \int \sum_j 2w_j \frac{\mathbf{1}\left(-e^{-\lambda_j} v < y < e^{\lambda_j} v\right) \cdot \mathbf{1}\left(0 < v < a_j\right)}{2\cosh\left(\lambda_j\right) \cdot a_j^2} dv$$

$$= \sum_j w_j \int_0^{a_j} \frac{\mathbf{1}\left(-e^{-\lambda_j} v < y < e^{\lambda_j} v\right)}{\cosh\left(\lambda_j\right) \cdot a_j^2} dv = \sum_j w_j \int_{\max\left(ye^{-\lambda_j}, -ye^{\lambda_j}\right)}^{a_j} \frac{1}{\cosh\left(\lambda_j\right) \cdot a_j^2} dv$$

$$\propto \sum_j w_j \frac{a_j - \max\left\{ye^{-\lambda_j}, -ye^{\lambda_j}\right\}}{\cosh\left(\lambda_j\right) \cdot a_j^2}.$$

As noted earlier, we would want to reduce the infinite mixture to a finite number of components,

similar to Walker (2007). To this end, introduce another auxiliary variable u and consider the joint density,

$$f(y,v,u) = \sum_j \mathbf{1}\left(u < w_j\right) \frac{\mathbf{1}\left(-e^{-\lambda_j} v < y < e^{\lambda_j} v\right) \mathbf{1}\left(0 < v < a_j\right)}{\cosh\left(\lambda_j\right) \cdot a_j^2} \tag{2.8}$$

Since $\sum w_j = 1$, it is clear that given u the number of components is finite, where the set of indices, $A_u = \{j : w_j > u\}$. The only issue left is to determine which of these finite components provides each datum. To accomplish this, introduce a final auxiliary variable, d, and consider the joint density

$$f(y,v,u,d) = \mathbf{1}\left(u < w_d\right) \frac{\mathbf{1}\left(-e^{-\lambda_d} v < y < e^{\lambda_d} v\right) \mathbf{1}\left(0 < v < a_d\right)}{\cosh\left(\lambda_d\right) \cdot \left(e^{-\lambda_d} + e^{\lambda_d}\right) \cdot a_d^2}.$$

The role of u is important since without it d could take on an infinite number of values, leading to a poor MCMC algorithm. Now, we have the complete likelihood function given by

$$L(a,\lambda,d,\mathbf{v},u|\mathbf{Y}) \propto \prod_{i=1}^{n} \mathbf{1}\left(u < w_{d_i}\right) \frac{\mathbf{1}\left(-e^{-\lambda_{d_i}} v_i < y_i < e^{\lambda_{d_i}} v_i\right) \mathbf{1}\left(0 < v_i < a_{d_i}\right)}{\cosh\left(\lambda_{d_i}\right) \cdot a_{d_i}^2} \tag{2.9}$$

To avoid simulation problems such as an increasing number of w's, one can apply a more general approach to slice sampling by introducing a deterministic, positive and decreasing sequence $\{\xi_1, \xi_2, \xi_3, ...\}$ into the model. Following Kalli et al. (2011), consider the generalized slicer sampler for the MDT model:

$$f(y,v,u,d) = \xi_d^{-1} \mathbf{1}\left(u < \xi_d\right) \cdot w_d \frac{\mathbf{1}\left(-e^{-\lambda_d} v < y < e^{\lambda_d} v\right) \cdot \mathbf{1}\left(0 < v < a_d\right)}{\cosh\left(\lambda_d\right) \cdot a_d^2}. \tag{2.10}$$

The choice of $\xi_1, \xi_2, \xi_3, ...$ is a delicate issue and any choice has to balance efficiency and computational time. Kalli et al. (2011) show that the mixing of the Markov chain depends on the rate at which the ratio $r_i \propto E[w_i]/\xi_i$ increases with i. Faster rates of increase are associated with better mixing but longer running times since the average size of A_u increases.

Using the ξ sequence, the new joint likelihood function is given by:

$$L(a,\lambda,d,\mathbf{v}|\mathbf{Y},\,\xi) = \prod_{i=1}^{n} \xi_{d_i}^{-1} \mathbf{1}\left(u_i < \xi_{d_i}\right) \cdot w_{d_i} \frac{\mathbf{1}\left(-e^{-\lambda_{d_i}} v_i < y_i < e^{\lambda_{d_i}} v_i\right) \mathbf{1}\left(0 < v_i < a_{d_i}\right)}{\cosh\left(\lambda_{d_i}\right) \cdot a_{d_i}^2} \tag{2.11}$$

It is now straightforward to write down the full conditional distributions needed to implement a

Gibbs sampler and to which we now turn. With p_0 denoting a prior distribution:

$$\pi\left(\lambda_j|...\right) \propto p_0(\lambda_j) \cdot \left(\frac{1}{\cosh\left(\lambda_j\right)}\right)^{k_j} \mathbf{1}\left\{\max\left(\ln\left(\frac{y_i^+}{v_i}\right) < \lambda_j < \min\left(\ln\left(\frac{-v_i}{y_i^-}\right)\right)\right)\right\} \ \forall i: \ d_i = j$$

where k_j is $\#\, d_i = j$.

$$\pi\left(a_j|...\right) \propto p_0\left(a_j\right) \cdot \prod_{d_i=j} \frac{1}{a_j^2} \mathbf{1}\left(0 < v_i < a_j\right) \propto p_0\left(a_j\right) \cdot \left(\frac{1}{a^2}\right)^{k_j} \mathbf{1}\left(a_j > \max v_i\right) \ \forall i: \ d_i = j$$

$$\pi\left(v_i|...\right) \propto \mathbf{1}\left(\max\left(-y_i e^{\lambda_{d_i}},\, y_i e^{-\lambda_{d_i}}\right) < v_i < a_{d_i}\right)$$

$$\pi\left(d_i = j \,|\, y_i, v_i, a_j, \lambda_j\right) \propto \xi_{d_i}^{-1} \mathbf{1}\left(u_i < \xi_j\right) \cdot w_{d_i} \frac{\mathbf{1}\left(-e^{-\lambda_j} v_i < y_i < e^{\lambda_j} v_i\right) \mathbf{1}\left(0 < v_i < a_j\right)}{\cosh\left(\lambda_j\right) \cdot a_j^2}$$

The priors for a_j and λ_j are taken to be Uniform and $N(\mu_0, \sigma_0)$, respectively. Armed with values for the hyperparameters, the Gibbs steps to implement an MTD model now follows.

In the simulated example section, details on sampling each of the above are described. Here it is simply noted that almost all the elements are fairly straightforward to deal with except for the sampling of λ. In section 2.1.2, we proposed using a Metropolis-Hastings (MH) algorithm in the simple one component triangle model. However, MH is not feasible here since in each Gibbs iteration, we are sampling different components of λ_j. We cannot keep track of the previous sample since they do not come from the same target distribution. The following alternative algorithm is proposed instead.

Adaptive rejection sampling (ARS) is a method for efficiently sampling from any univariate probability density function which is log-concave; see Gilks and Wild (1992). If we take a normal prior for $\lambda_j \sim N(0,\, \sigma^2)$, then the kernel of the posterior distribution of λ_j is

$$\pi\left(\lambda_j|...\right) \propto \left(e^{-\frac{\lambda_j^2}{2\sigma^2}}\right) \left(\frac{1}{\cosh\left(\lambda_j\right)}\right)^{k_j}.$$

Taking logs,

$$\log\left(\pi(\lambda_j|...)\right) \propto -\frac{\lambda_j^2}{2\sigma^2} - k_j \log\left(\cosh\left(\lambda_j\right)\right),$$

Algorithm 1 Gibbs Sampler for MTD

1. Set initial value of d_i (preferably some small value) and feasible initial value of (v_i).
2. Sample u_i given d_i and ξ_{d_i} where $\pi\left(u_i|...\right) \propto \mathbf{1}\left(0 < u_i < \xi_{d_i}\right)$.
3. Sample δ_j given d_i, where the number of components can be calculated by u_i and ξ_{d_i}; and so, $\pi\left(\delta_j|...\right) \propto \text{Be}(\delta_j : \alpha_j,\, \beta_j)$ where $\alpha_j = 1 + \sum \mathbf{1}\left(d_i = j\right)$ and $\beta_j = M + \sum \mathbf{1}\left(d_i > j\right)$.
4. Calculate weights w_j given δ_j.
5. Sample a_j from $\pi\left(a_j|...\right)$ and λ_j from $\pi\left(\lambda_j|...\right)$.
6. Sample v_i from $\pi\left(v_i|...\right)$.
7. Sample d_i as follows: $\text{P}\left(d_i = k\,|...\right) = \mathbf{1}\left(k\,:\,\xi_k > u_i\right)\cdot\omega_k/\xi_k\cdot\text{Tri}\left(y_i;\, a_k,\lambda_k\right)$ where $\text{Tri}\left(y_i;\, a_k,\lambda_k\right) = \frac{a_k - \max\left\{y_i e^{-\lambda_k}, -y_i e^{\lambda_k}\right\}}{\cosh(\lambda_k)\cdot a_k^2}$.
8. Repeat steps 2 - 7.

---

from which the first and second derivatives are obtained as,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\log\left(\pi(\lambda_j|...)\right) \propto -\frac{\lambda_j}{\sigma^2} - k_j\left(\frac{\sinh\left(\lambda_j\right)}{\cosh\left(\lambda_j\right)}\right) = -\frac{\lambda_j}{\sigma^2} - k_j\cdot\left(\tanh\left(\lambda_j\right)\right)$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\log\left(\pi(\lambda_j|...)\right) \propto -\frac{1}{\sigma^2} - k_j\cdot\operatorname{sech}^2\left(\lambda_j\right) < 0,$$

establishing log concavity of the posterior distribution of λ_j.

Since we assume an infinite number of components in the mixture, we do not really need to keep track of all the parameters in each iteration. However, we can predict the next observation by first sampling a weight indicator variable η from a uniform(0,1) distribution. Then we can line up the w_j's and check which component this new observation falls into. Finally, we can sample the new observation given the parameters of the sampled component. Note that in this algorithm, one cannot sample all the components. As a result, if our weight indicator falls outside all the components, for example if $\eta > \sum w_j$, then we need to first sample the component from the prior distribution of the parameters and use those parameters to sample a new observation.

3 Simulation Experiments for the Triangular and MTD Models

In this chapter, simulated data illustrations are conducted to exemplify the models developed in Chapter 2.

3.1 Simulations for the Triangular Distribution

Observations from symmetric and asymmetric distributions are used to illustrate Bayesian estimation using the triangular distributions developed in Chapter 2. Specifically, the following distributions are used.

- Triangle($a = 3, \lambda = 0.5$)
- Triangle($a = 4, \lambda = -1$)
- Normal$(0, 1)$
- Beta$(3, 2)$ where the density was shifted so that its mode equals zero.

As discussed in Chapter 2, in all the simulation examples, the conditional distribution of $(\lambda|...)$ is the only non-standard distribution in the Gibbs sampler that poses difficulty. To sample this, using a Gaussian proposal, an independent M-H algorithm was implemented.

The acceptance probability α takes the form:

$$\alpha = \max\left(1, \frac{Q(x_t)P(x')}{Q(x')P(x_t)}\right)$$

where $Q(\cdot)$ is the proposal density and $P(\cdot)$ is the target density. In sampling λ, with a Gaussian prior $(0, \sigma_\lambda)$, a truncated Gaussian density with mean zero and standard deviation σ_p is used as the proposal density. With x' and x_t denoting draws from the proposal density and the current value of x, respectively, the acceptance probability is given by:

$$\log\left(\frac{Q(x_t)P(x')}{Q(x')P(x_t)}\right) = -n \cdot \left((e^{-x_t} + e^{x_t}) - (e^{-x'} + e^{x'})\right) - \frac{1}{2\sigma_\lambda}(x_t^2 - x'^2) - \frac{1}{2\sigma_p}(x'^2 - x_t^2)$$

Table 1 summarizes the details of the simulations for each of the four models.

Table 1: Simulation Settings

True Densities	# of observations (M)	# of iterations (N)	Burn-in	Thinning
Triangular (3, 0.5)	500	100,000	5000	50
Triangular (4, -1)	500	100,000	5000	50
Normal (0, 1)	500	30,000	5000	10
Beta (3, 2) at mode 0	500	100,000	5000	50

Table 2: Summary Statistics of the Posterior Distributions

	Triangle(3, 0.5)	Triangle(4,-1)	Normal(0,1)	Beta(3,2)
a				
True Value	3.000	4.000	NA	NA
Mean	3.106	4.084	3.135	0.470
SD	0.066	0.129	0.025	0.009
95% Interval	(3.004, 3.259)	(3.863, 4.323)	(3.099, 3.196)	(0.455, 0.489)
λ				
True Value	0.500	-1.000	NA	NA
Mean	0.497	-0.987	0.002	-0.251
SD	0.019	0.030	0.008	0.018
95% Interval	(0.457, 0.533)	(-1.032, -0.913)	(-0.012, 0.018)	(-0.285, -0.211)

Consider Table 2 where the true values, the posterior means, standard deviations and 95% posterior credibility intervals for a and λ are given. There is close agreement between the estimates and the true values in the case of the two triangular densities for which we know the exact values of a and λ. For the normal distribution the skewness parameter is zero and so any estimate of λ should be close to zero as shown in Table 2. Also, the kurtosis parameter a is close to 3 which is the kurtosis of a normal distribution. The Beta$(3, 2)$ is skewed left and this is evidenced by the fact that the posterior estimates for λ are negative. Note that we cannot compare the a estimate to the kurtosis of the Beta$(3, 2)$ density since the triangular and beta densities are scaled differently.

Consider Figures 1, 2, 3 and 4 that contain overlaid plots of the true density (solid curve), density estimate using the $M = 500$ samples (dotted curve) and the MCMC-based estimate of the true density (bold dotted curve). As expected, the kurtosis parameter, a, forces the latter estimates to have fatter tails. The Bayesian triangular distribution estimates also capture the asymmetry of the Triangle$(3, 0.5)$, Triangle$(4, -1.0)$, and Beta$(3, 2)$ densities very well.

Figure 1: The Triangular Density Estimation for Triangular (3, 0.5)

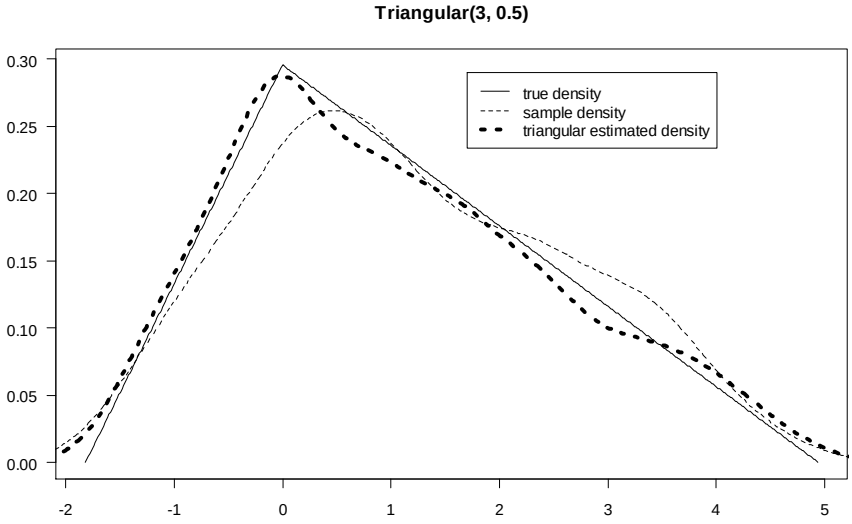


Figure 2: The Triangular Density Estimation for Triangular (4, -1)

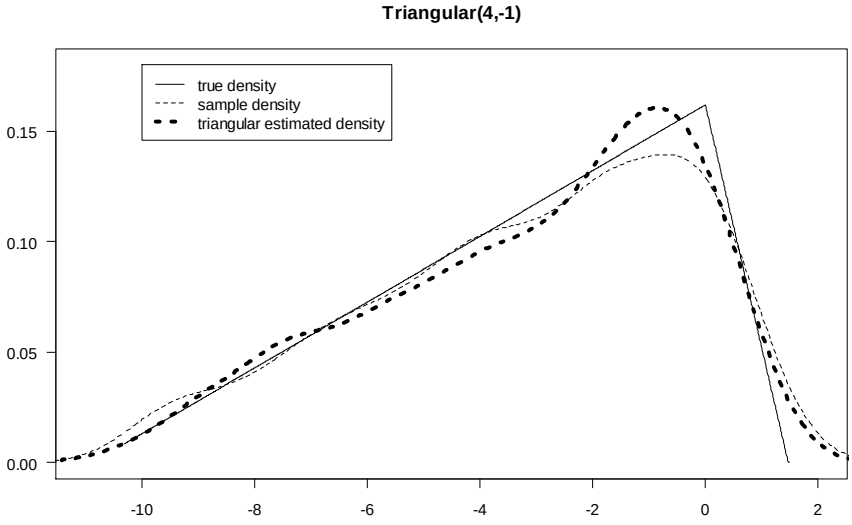


Figure 3: The Triangular Density Estimation for Gaussian(0, 1)

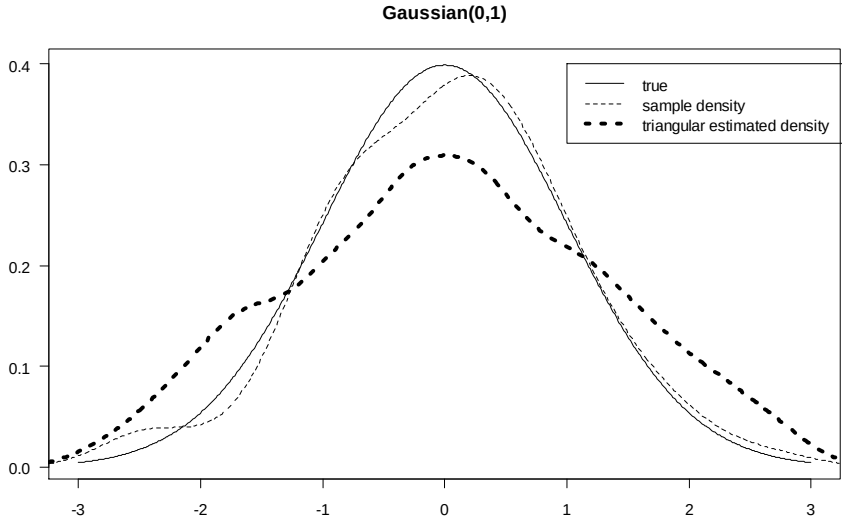
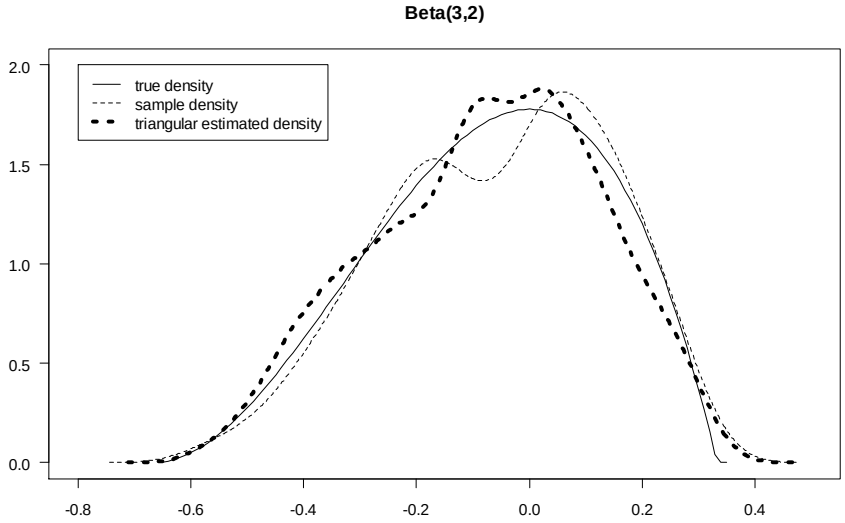


Figure 4: The Triangular Density Estimation for Beta(3, 2)



3.2 Simulations for the MTD Model

Here the following three densities were selected to study the performance of the MTD model: (a)

Normal$(0, 1)$; (b) Laplace$(0, 4)$; and (c) Chi-squared(5) with mode at zero.

The priors for a and λ are the same as the ones used in the previous section. The prior distribution for the weights, w_i, are taken to be

$$\delta_i \sim \text{Beta}(1,\ 4).$$

The ξ_i values are set to be $(0.1667)^{-i}$.

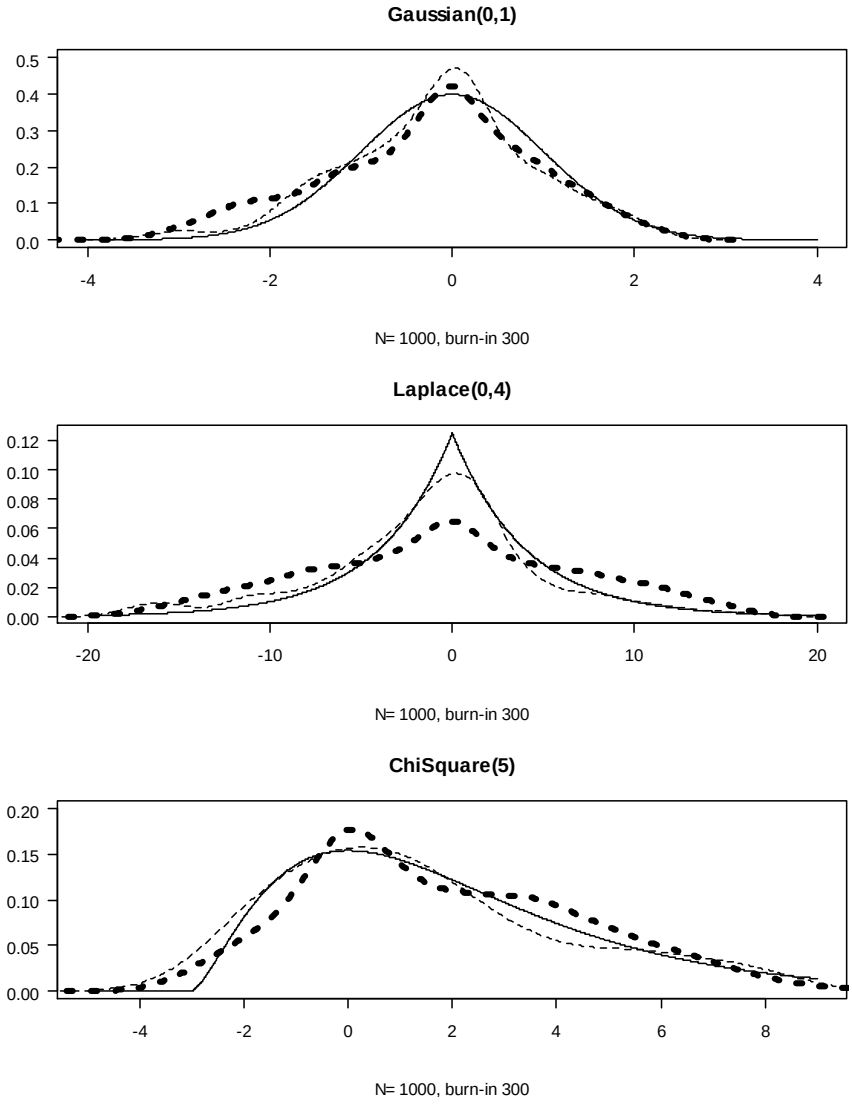
With $M = 50$ observations, $N = 1000$ iterations and 300 burn-in samples, Table 3 provides different percentile values from the true density and the MTD-based Bayesian estimates; there is good agreement between these values.

The three panels in Figure 5 illustrate the posterior density estimates using the MTD model. The solid curve in Figure 5 is the true density while the dotted curve and the bold dotted curve are density estimates using the $M = 50$ samples drawn from the true density and the MCMC-based Bayesian MTD estimates, respectively.

Table 3: The Percentiles of the True Densities and the MTD Density Estimates

	Gaussian(0,1)		Laplace(0,4)		Chisquare(5) at mode 0	
	True	MTD	True	MTD	True	MTD
1%	-2.33	-2.96	-15.65	-15.77	-2.45	-2.86
5%	-1.64	-2.39	-9.21	-12.60	-1.85	-1.92
25%	-0.67	-1.02	-2.77	-5.45	-0.33	-0.06
50%	0.00	-0.08	0.00	-0.05	1.35	1.26
75%	0.67	0.55	2.77	4.77	3.63	3.64
95%	1.64	1.56	9.21	11.88	8.07	6.53
99%	2.33	2.13	15.65	14.84	12.09	8.09

Figure 5: The MTD Estimates for the Three Underlying Densities



4 Bayesian Mode Regressions

Full Bayesian implementation of mode regressions using triangular and MTD errors is developed in this chapter. As discussed in the Introduction, the major advantage of using a triangular error term in mode regression is to be able to model the data distribution when it is unimodal, symmetric or asymmetric, and convex with varying levels of kurtosis to account for fat tail behavior. Such distributional forms could also arise when the observed data are elicitations of the minimum, mode, and maximum values from experts in applied contexts such as petroleum exploration.

4.1 Mode Regression with Triangular Error

Consider the following regression model:

$$y = X\beta + \varepsilon \,\text{where}\, \varepsilon \sim Triangular(a, \lambda),$$

that is, the distribution of ε is triangular with mode at zero. The new wrinkle in implementing this regression model is to be able sample from the posterior distribution of the regression parameter, β. The structure of the conditional distributions for all the other unknowns is similar to the one described in Chapter 2. Thus, using the latent form of the model from section 2.1.3, and with $\pi(.)$ denoting a prior distribution, it is easy to obtain the following full conditional distributions.

$$f(v_i|y_i, a, \lambda, \beta) \propto \mathbf{1}\left(\max\left\{(y_i - \beta X_i)e^{-\lambda}, -(y_i - \beta X_i)e^{\lambda}\right\}, a\right)$$

$$f(a|\mathbf{v}, y, \lambda) \propto \pi(a) \cdot \prod_i f(v_i|a) \propto \pi(a) \cdot \frac{1}{a^{2n}} \mathbf{1}(a > \max v_i)$$

$$f(\lambda|y, \mathbf{v}, a, \beta) \propto \pi(\lambda) \cdot \prod_i f\left(y_i - \beta X_i|v_i, \lambda\right)$$

$$\propto \pi(\lambda) \cdot \left(\frac{1}{(\exp(-\lambda) + \exp(\lambda)}\right)^n \mathbf{1}\left(\max\left(\ln(\frac{(y_i - \beta X_i)^+}{v_i})\right) < \lambda < \min\left(\ln(\frac{-v_i}{(y_i - \beta X_i)^-})\right)\right)$$

For a simple linear (mode) regression model, the likelihood function for β is

$$L(\beta|...) = \prod_i f(y_i - \beta X_i|a,\, \lambda,\, v_i) \propto$$

$$\propto \prod \mathbf{1}\left(\frac{y_i - a\lambda}{x_i} < \beta < \frac{y_i + a\lambda^{-1}}{x_i} |\ x_i > 0\right) \cdot \mathbf{1}\left(\frac{y_i + a\lambda^{-1}}{x_i} < \beta < \frac{y_i - a\lambda}{x_i} |\ x_i < 0\right)$$

Thus, the likelihood function for β is flat and the shape of the posterior distribution of β is completely dependent on the prior distribution of β, $\pi(\beta)$. One reasonable choice for the prior is a normal distribution, since then the posterior distribution is simply a truncated normal; hence sampling from the posterior conditional distribution for β is straightforward. Also note from the form of the likelihood function, the truncation limits will get smaller as sample size increases. This point is critical when we transit from a simple to a multiple regression model.

In the case of a multiple linear (mode) regression model, sampling a, λ, and v_i remains unchanged. However, sampling the parameter vector β becomes challenging. Consider the case with just two regressors and four observations. The boundary constraints for the coefficients are obtained as

$$\begin{aligned} a_1 &< \beta_1 x_{11} + \beta_2 x_{12} < b_1 \\ a_2 &< \beta_1 x_{21} + \beta_2 x_{22} < b_2 \\ a_3 &< \beta_1 x_{31} + \beta_2 x_{32} < b_3 \\ a_4 &< \beta_1 x_{41} + \beta_2 x_{42} < b_4. \end{aligned}$$

The support region for the posterior joint distribution of (β_1, β_2) is a convex polygon. Given the data, we can solve all the pairwise equations to get the exact support region; however, it can be very time-consuming. As a result, we will model the prior distributions of the coefficients to be independent normal distributions. This reduces the problem to solving a one dimensional boundary condition whereby the elements of β are sampled successively, starting with β_1.

Another practical issue is to come up the initial values to start the Markov chain. We first use the OLS estimates of the coefficients to calculate ε_i's to ensure $\varepsilon_{min} < 0 < \varepsilon_{max}$. Then, use ε_{min} and ε_{max} to estimate a and λ as

$$a_{est} = (\varepsilon_{max} \cdot \varepsilon_{min})^{1/2}$$

$$\lambda_{est} = \frac{1}{2} \cdot \log(\varepsilon_{max}/\varepsilon_{min}).$$

Now, to start the Markov chain, set the initial value a_{init} to be slightly larger than a_{est} and let $\lambda_{init} = \lambda_{est}$. In the case of symmetric errors, set $\lambda_{init} = 0$ and $a_{init} > \max(\varepsilon_{max}, -\varepsilon_{min})$. Regardless of the symmetry assumption, the variables that need to be sampled at each sweep of a Gibbs sampler

are $\{(\beta_1, ..., \beta_m, a, \lambda, v_1, ...v_n\}$ where m is the number of regressors.

The above model will be exemplified using simulated data in the following chapter.

4.2 Mode Regressions with MTD Errors

With the Gibbs sampler of MTD in hand, we now consider the following regression model

$$y = X\beta + \varepsilon \text{ where } \varepsilon \sim MTD,$$

that is, the distribution of ε's are MTD with zero modes for all the components.

Previously, we developed the joint likelihood for the MTD given data and ξ sequence in equation (3.5). We now replace y with the error terms ε and the joint likelihood for ε becomes:

$$L(a, \lambda, d, \mathbf{v}|\mathbf{Y}, X, \beta\, \xi) = \prod_{i=1}^{n} \xi_{d_i}^{-1} \mathbf{1}\left(u_i < \xi_{d_i}\right) \cdot w_{d_i} \frac{\mathbf{1}\left(-e^{-\lambda_{d_i}} v_i < \varepsilon_i < e^{\lambda_{d_i}} v_i\right) \mathbf{1}\left(0 < v_i < a_{d_i}\right)}{\cosh\left(\lambda_{d_i}\right) \cdot a_{d_i}^2} \tag{4.1}$$

The Gibbs samplers for a, λ, d and $\mathbf{v}$ are the same as in section 2.2.3 given the regression coefficients β.

Since we will be sampling each regression coefficient given all others, consider the model with one regressor:

$$y_i = \beta x_i + \varepsilon_i$$

The likelihood function for β given all the other parameters is

$$L(\beta|...) = \prod_i f\left(y_i - \beta x_i | a, \lambda, v_i\right)$$

$$= \prod_i \xi_{d_i}^{-1} \mathbf{1}(u_i < \xi_{d_i}) \cdot w_{d_i} \frac{\mathbf{1}\left(-e^{-\lambda_{d_i}} v_i < y_i - \beta x_i < e^{\lambda_{d_i}} v_i\right) \mathbf{1}\left(0 < v_i < a_{d_i}\right)}{\cosh\left(\lambda_{d_i}\right) \cdot a_{d_i}^2}$$

$$\propto \prod_i \mathbf{1}\left(-e^{-\lambda_{d_i}} v_i < y_i - \beta x_i < e^{\lambda_{d_i}} v_i\right)$$

$$\propto \prod_i \mathbf{1}\left(\frac{y_i - v_i\lambda_{d_i}}{x_i} < \beta < \frac{y_i + v_i\lambda_{d_i}^{-1}}{x_i}\middle| x_i > 0\right) \cdot \mathbf{1}\left(\frac{y_i + v_i\lambda_{d_i}^{-1}}{x_i} < \beta < \frac{y_i - v_i\lambda_{d_i}}{x_i}\middle| x_i < 0\right)$$

Since $n > 0$, β is always bounded. Using a uniform prior, the conditional posterior distribution of each β should be uniform as well.

$$\pi(\beta|v, \lambda ...) \propto \prod_i \mathbf{1}\left(\frac{y_i - v_i\lambda_{d_i}}{x_i} < \beta < \frac{y_i + v_i\lambda_{d_i}^{-1}}{x_i}\middle| x_i > 0\right) \cdot \mathbf{1}\left(\frac{y_i + v_i\lambda_{d_i}^{-1}}{x_i} < \beta < \frac{y_i - v_i\lambda_{d_i}}{x_i}\middle| x_i < 0\right)$$

The priors for a_j, λ_j are again taken to be Pareto(0.05, 3) and $N(0, \sigma_0)$ respectively. The Gibbs steps to implement an MTD regression model are very similar to Gibbs steps for MTD alone. The only difference is that one extra step is needed to sample β.

To expand the above algorithm to have multiple regressors, assume independent priors for each β and add a block in step 7 to sample β_k given the rest of the βs and other parameters. Due to the nature of truncations in the conditional posterior distributions, sampling β at once as a truncated multivariate normal is not advisable.

The predictive distribution of y_i can be sampled by taking an extra step in the MCMC algorithm. We first sample a uniform index variable η and use w_j to determine which components y_i comes from. It either comes from the components we have in the Gibbs iteration or outside the components we have sampled which means$\eta > \sum w_j$. For the former case, use the corresponding (a_j, λ_j) to sample the residuals and for the latter, sample (a, λ) from their priors to sample the residuals. Then, we can calculate the predictive value of y_i given the value of β and ε in each iteration.

4.3 Consistency of Mode Regression with MTD Errors

In this section, we investigate the consistency properties of the modal regression model developed earlier. To this end, we first consider consistency of just the non-regression model which could be dealt using the idea from Walker and Hjort (2001). In the following, we are dealing with f being a unimodal density with mode at zero. The posterior assigned to the set A is given by

$$\Pi_n(A) = \Pi(A|Y_1, \ldots, Y_n) = \frac{\int_A R_n(f)\,\pi(df)}{\int R_n(f)\,\pi(df)},$$

Algorithm 2 Gibbs Sampler for Mode Regression with MTD Errors

1. Set Ordinary Least Square (OLS) estimator to be the initial value of β and calculate the residuals ε.

2. Set initial value of d_i (preferably some small number) and feasible initial value of v_i= max(ε_i,-ε_i) .

3. Sample u_i given d_i and $\xi_{d_i}$where $\pi(u_i|...) \propto \mathbf{1}(0 < u_i < \xi_{d_i})$.

4. Sample δ_j given d_i, where the number of components can be calculated by u_i and ξ_{d_i}; and so, $\pi(\delta_j|...) \propto \text{Be}(\delta_j : \alpha_j, \beta_j)$ where $\alpha_j = 1 + \sum \mathbf{1}(d_i = j)$ and $\beta_j = M + \sum \mathbf{1}(d_i > j)$.

5. Calculate weights w_j given δ_j.

6. Sample a_j from $\pi(a_j|...)$ and λ_j from $\pi(\lambda_j|...)$.

7. Sample β given λ_j, d_i and ε_i

8. calculate the new residuals ε_i.

9. Sample v_i from $\pi(v_i|...)$.

10. Sample d_i as follows: $\text{P}(d_i = k|...) = \mathbf{1}(k : \xi_k > u_i)\omega_k/\xi_k \cdot \text{Tri}(y_i; a_k, \lambda_k)$ where $\text{Tri}(y_i; a_k, \lambda_k) = \frac{a_k - \max\{y_i e^{-\lambda_k}, -y_i e^{\lambda_k}\}}{\cosh(\lambda_k) \cdot a_k^2}$.

where $R_n(f) = \prod_{i=1}^n f(Y_i)/f_0(Y_i)$. Write the numerator L_n as

$$L_n \leq \left(\prod_{i=1}^n \frac{\widehat{f}}{f_0}(Y_i) \right)^{n/2} \int_{A_\epsilon} R_n^{1/2}(f)\, \pi(df)$$

where $\widehat{f}$ is the non-parametric MLE and

$$A_\epsilon = \{f : H(f_0, f) > \epsilon\}$$

and H is the Hellinger distance.

Using standard results (see Walker and Hjort, 2001),

$$\int_{A_\epsilon} R_n^{1/2}(f)\, \pi(df) \leq e^{-n\delta(\epsilon)} \quad \text{a.s.}$$

for all large n, for some $\delta(\epsilon) > 0$.

From van de Geer (1993), knowing that f is unimodal with mode at 0, it is that

$$n^{-1} \sum_{i=1}^n \log\{\widehat{f}(Y_i)/f_0(Y_i)\} \to 0 \quad \text{a.s.}$$

and hence

$$L_n \leq e^{-n\tilde{\delta}(\epsilon)} \quad \text{a.s.}$$

for all large n for some $\tilde{\delta}(\epsilon) > 0$.

This is now sufficient for showing that

$$\Pi_n(A_\epsilon) \to 0 \quad \text{a.s.}$$

for all $\epsilon > 0$ for all f_0 in the Kullback-Leibler support of the prior. This establishes posterior consistency w.r.t. the nonparametric component for the class of densities developed in this paper.

Now consider the regression model,

$$y_i = \alpha + \beta\, x_i + \epsilon_i$$

with the (ϵ_i) i.i.d. from f which is unimodal at 0 and convex either side of 0.

The work for this type of model has been done by Amewou-Atisso et al. (2003). Specifically, we would like to apply the following theorem from their paper.

Theorem (Amewou-Atisso et al. (2003)): *Suppose* $\tilde{\Pi}$ *is a prior on* $\mathcal{F}$ *and* μ*is a prior for* (α, β). *Let* $\mathcal{W} \subset \mathcal{F} \times \mathbb{R} \times \mathbb{R}$. *If*

(i) there is an exponential consistent sequence of tests for

$$H_0 : (f, \alpha, \beta) = (f_0, \alpha_0, \beta_0) \quad \text{against} \quad H_1 : (f, \alpha, \beta) \in \mathcal{W},$$

(ii) and for all $\delta > 0$,

$$\Pi \left\{ (f, \alpha, \beta) : K_i(f, \alpha, \beta) < \delta \, for \, all \, i, \quad \sum_{i=1}^{\infty} \frac{V_i(f, \alpha, \beta)}{i^2} < \infty \right\} > 0$$

then with $(\prod_{i=1}^{\infty} P_{f_{0i}})$ *– probability* 1, *the posterior probability*

$$\Pi \left(\mathcal{W} | Y_1, \ldots, Y_n \right) = \frac{\int_{\mathcal{W}} \prod_{i=1}^{n} \left(f_{\alpha,\beta i}(Y_i) / f_{0i}(Y_i) \mathrm{d} \prod(f, \alpha, \beta) \right)}{\int_{\mathcal{F} \times \mathbb{R} \times \mathbb{R}} \prod_{i=1}^{n} \left(f_{\alpha,\beta i}(Y_i) / f_{0i}(Y_i) \mathrm{d} \prod(f, \alpha, \beta) \right)} \to 0$$

The consistency of the posterior holds as long as there is an exponentially consistent test for testing the point null against the complement of the required neighborhood and (ii) holds. Following Assumptions A, B and Propositions 3.1, 3.2, and 3.3 from Amewou-Atisso et al. (2003), such exponentially consistent test exists for our MTD regression models.

The authors assume f to be symmetric about 0 in order for the pair (α, f) to be identifiable, in order to appply the above thoerem. We claim that identifiability, and hence consistency for (α, β), also follows if f has mode at 0 and f is convex either side of 0.

To see this, let us consider one side of 0 and ask if there can be (g, f, α, α') such that

$$g(y - \alpha) \equiv f(y - \alpha').$$

So here $g(y)$ and $f(y)$ would be convex and decreasing densities on $(0, \infty)$. From the convexity of f we have

$$f(y - \alpha') \geq f(y - \alpha) + f'(y - \alpha) \, (\alpha - \alpha').$$

Since $f' < 0$ and we take $\alpha' > \alpha$ without loss of generality, we have

$$f(y - \alpha') = g(y - \alpha) > f(y - \alpha)$$

which is not possible since f and g are density functions. Hence, (f, α) is identifiable under the condition of convexity.

To deal with the two-sided case, we need to show that if

$$f_\pi(y - \alpha) = \pi f_-(y - \alpha) + (1 - \pi)\, f_+(y - \alpha)$$

is equivalent to

$$f_{\pi'}(y - \alpha') = \pi' f_-(y - \alpha') + (1 - \pi')\, f_+(y - \alpha'),$$

where f_- and f_+ are the normalized negative and positive parts of f, respectively, and

$$\pi = \int_{y<\alpha} f_-(y - \alpha)\, dy,$$

then $(\alpha, \pi) = (\alpha', \pi')$. However, this follows trivially since

$$\int_{y<\alpha} f_-(y - \alpha)\, dy$$

does not depend on α.

Having established identifiability, we can now invoke the above thoerem from Amewou-Atisso et al. (2003), to establish consistency for the parametric component of the family of MTD-based regression models.

5 Simulated Experiments for Bayesian Mode Regressions

In this chapter, simulated data illustrations are conducted to exemplify the models developed in Chapter 4.

5.1 Simulations for Bayesian Mode Regression with Triangular Error

The model is set up as

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \text{ where } \varepsilon \sim \text{Triangular}(a, \lambda).$$

Setting $\beta_1 = 2$, $\beta_2 = 2$, $a = 3$, and $\lambda = 1$, we obtain y values after simulating $M = 500$ observations using $x_1 \sim Normal(0, 3)$, $x_2 \sim Normal(0, 1)$ and ε. Given these observations, Bayesian mode regression, via the model in Section 4.1, is performed. To this end, let the prior distributions of $\{\beta_1, \beta_2, a, \lambda\}$ be:

$$p(\beta_1) \sim \text{Normal}(0, 10); \quad p(\beta_2) \sim \text{Normal}(0, 10)$$

$$p(a) \sim \text{Pareto}(0.1, 2); \quad p(\lambda) \sim \text{Normal}(0, 1).$$

The initial values for the regression coefficients, b_1 and b_2, are calculated via OLS estimation; these, in turn, are used to obtain the errors $\hat{\varepsilon} = y - b_1 x_1 + b_2 x_2$.

The initial values of a, λ are $\max(\max(\hat{\varepsilon}), -\min(\hat{\varepsilon}))$, and 0, respectively.

Some key summary statistics in Table 4 for the posterior distributions of β_1, β_2, a, and λ show good approximation to the true values of these parameters.

Consider Figure 6 that depicts the predictive distribution of y obtained using $x_1 = 1$ and $x_2 = 2$. A 95% interval encapsulating the true value of y is shown in this graph. Figure 7 shows the distribution of ϵ whose theoretical mode is zero; as expected, this graph is asymmetric since it is consistent with

Table 4: Posterior Summary Statistics for Mode Regression Example

	True Value	Posterior Mean	Posterior SD	Posterior 95% Interval
a	3	3.198	0.119	(2.997, 3.504)
λ	1	0.913	0.036	(0.837, 0.979)
β_1	2	2.003	0.051	(1.919, 2.101)
β_2	3	2.908	0.151	(2.642, 3.169)

the assumption that the distribution of the error is skewed with $\lambda = 1$.

Figure 6: The Predictive Density of y

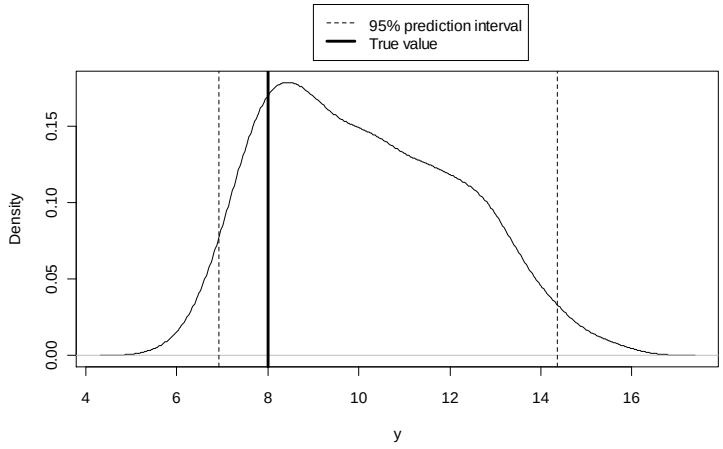
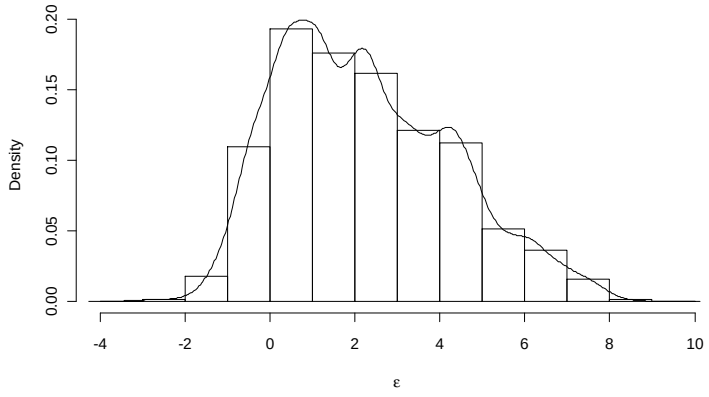


Figure 7: The Distribution of ε



5.2 Simulations for Bayesian Mode Regression with MTD Error

In this subsection, the Bayesian MTD mode regression model is illustrated via simulated data.

Table 5: The Parameters and Density Plots for Simulation Errors.

		1st Component	2nd Component	3rd Component	Density Plot
Mix 1	p	0.50	0.3	0.2	
	min	-0.05	-2.0	-4.0	
	max	0.05	5.0	4.0	
Mix 2	p	0.30	0.3	0.4	
	min	-1.00	-2.0	-3.0	
	max	1.00	5.0	2.0	

Table 6: Simulation Example 1: True Parameter Values (TV) and the Corresponding Posterior Means, Standard Deviation (SD), 95% Credible Intervals (CI) and the OLS Estimator as Initial Value (I.V)

n		Mix 1		Mix 2	
		β_0	β_1	β_0	β_1
50	T.V	0.000	1.000	0.000	1.000
	Mean	-0.002	1.004	0.012	1.020
	S.D.	0.011	0.009	0.081	0.078
	95% HPD	(-0.023, 0.020)	(0.985, 1.020)	(-0.136, 0.165)	(0.905, 1.189)
	I.V	0.220	1.111	0.426	1.094
100	T.V	0.000	1.000	0.000	1.000
	Mean	0.001	1.000	0.137	0.969
	S.D.	0.007	0.007	0.107	0.113
	95% HPD	(-0.012, 0.016)	(0.987, 1.016)	(-0.07, 0.395)	(0.827, 1.221)
	I.V	0.070	0.961	0.220	1.046
200	T.V	0.000	1.000	0.000	1.000
	Mean	-0.007	0.997	0.162	1.084
	S.D.	0.007	0.006	0.185	0.095
	95% HPD	(-0.020, 0.007)	(0.985, 1.010)	(-0.253,0.422)	(0.900, 1.269)
	I.V	0.200	0.900	0.152	1.124

These were also studied in Yu and Aristodemu (2014). We begin our simulations by setting the regression errors to be a three-component mixture of triangular distributions.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $x_i \sim N(0,1)$ and $i = 1, ..., n$ and $f(\epsilon) = p_1 f(z_1) + p_2 f(z_2) + p_3 f(z_3)$, $p_1 + p_2 + p_3 = 1$. Each z_i follows a triangular distribution with mode 0. We simulate two cases with different kurtosis and skewness. The first case has one center spike with small noise at the tail. The second is a smooth mixture. Both cases are asymmetric. Table 5 summarizes the parameters of the error distributions. The density plots are estimated from 200 samples which we used in our simulations.

Table 6 summarizes the results for the simulation. It is evident that the MTD model recovers the true values $\beta_0 = 0$ and $\beta_1 = 1$ nicely. The credible intervals quantify the uncertainty in the parameter

estimates.

In order to facilitate a direct comparison with the algorithm proposed by Yu and Aristodemu (2014) the same simulated data models used by them are considered here. Let

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $x_i \sim N(0,1)$ and $i - 1, ..., n$ with n =50, 100, 200. We set $\boldsymbol{\beta} = (1, 2)$ and consider two different error distributions ϵ_i.

- Case 1: the standard normal distribution, $\epsilon_i \sim N(0,1)$ - a symmetric error distribution;
- Case 2: Fisher's Z distribution, $\epsilon_i \sim 1/2 \log(Z)$ with $Z \sim F_{2,2}$ - an asymmetric error distribution

Simulation Notes: The MTD mode regression model was estimated via MCMC that was run for 200,000 iterations. After appropriately adjusting for autocorrelation in the chain by burning-in and thinning the chain, a sample of $2,000$ resulted, which was used to calculate the posterior estimates. Convergence was assessed using ACF plots and Geweke scores. The initial value for the coefficients are set to be the OLS estimators.

Table 7 summarizes the results for Case 1. All the true values of the parameters are contained in their respective 95% credible intervals. Also, when compared to the Parametric Bayesian Mode Regression (PBMR) of Yu and Aristodemu (2013), ours have smaller credible intervals.

Table 8 summarizes the results for Case 2. Again, when compared to the PBMR, ours have much smaller credible intervals. This is because PBMR performs worse under asymmetric errors. Here, IV stands for initial value which we set to be the OLS estimator. For Case 1, since the errors are symmetric, the OLS estimator is also an unbiased estimator for mode regression. However, with asymmetric errors, OLS estimator is clearly biased. We report the initial values and show that our algorithm converges to the true coefficients.

Table 7: Simulation Example 2 Case 1: True Parameter Values (TV) and the Corresponding Posterior Means, Standard Deviation (SD) and 95% Credible Intervals (CI)

		MTD		PBMR	
n		β_0	β_1	β_0	β_1
50	TV	1.000	2.000	1.00	2.00
	Mean	1.043	1.964	0.92	2.00
	S.D.	0.139	0.101	0.78	0.77
	95% CI	(0.860, 1.363)	(1.824, 2.196)	(-0.6, 2.1)	(0.5, 3.3)
100	TV	1.000	2.000	1.00	2.00
	Mean	1.016	2.090	1.01	2.10
	S.D.	0.169	0.106	0.18	0.25
	95% CI	(0.677, 1.89)	(1.884, 2.277)	(0.6, 1.3)	(1.6,2.6)
200	TV	1.000	2.000	1.00	2.00
	Mean	1.095	2.029	1.26	0.99
	S.D.	0.149	0.080	0.86	0.52
	95% CI	(0.819 1.397)	(1.913 2.208)	(-0.5,2.8)	(0.9, 3.0)

Table 8: Simulation Example 2 Case 2: True Parameter Values (TV) and the Corresponding Posterior Means, Standard Deviation (SD) and 95% Credible Intervals (CI)

n	Case 2	MTD		PBMR	
		β_0	β_1	β_0	β_1
50	T.V	1	2	1	2
	I.V	1.110	2.041		
	Mean	0.976	2.038	1.07	2.01
	S.D.	0.125	0.076	0.78	0.49
	95% C.I	(0.683 1.123)	(1.867 2.148)	(-0.3, 2.6)	(1.2, 3.1)
100	T.V	1	2	1	2
	I.V	0.896	2.163		
	Mean	0.929	2.157	0.95	1.89
	S.D.	0.110	0.109	0.52	0.37
	95% C.I	(0.670 1.107)	(1.914 2.326)	(0.0, 1.9)	(1.2, 2.6)
200	T.V	1	2	1	2
	I.V	0.976	2.025		
	Mean	1.088	2.037	1.00	1.99
	S.D.	0.073	0.056	1.29	0.75
	95% C.I	(0.957 1.235)	(1.930 2.137)	(-1.3, 3.5)	(0.6, 3.3)

6 Real Data Illustrations

In this chapter, we illustrate our model with two sets of data. The first one is from a paper by Yu and Aristodemou (2014) who model the conditional mode of worker productivity at Western Electric Workers Company (WECO) given factors such as gender, pre-employment test result and education. In addition to providing full Bayesian analysis of the regression coefficients, we also assess the viability of the model via out-of-sample predictions. The second set of data is a cross-sectional study of End Stage Renal Disease (ESRD) adults using a nationally representative Medical Expenditure Panel Survey (MEPS)[1]. We will analyze the medical expenses of patients by different critera each with several covariates. We then compare our result with a mean regression model (using OLS) since a non-informative Bayesian mean regression produces results that are qualitatively similar.

6.1 Productivity of Western Electric Company (WECO) Workers

To illustrate the applicability of our model, we analyze the productivity of newly hired electric workers in a manufacturing firm. To compare our results to Yu and Aristodemu (2013), we used the same data and regression model where productivity for the ith person is denoted (y_i), gender is an indicator variable, (sex_i), the score on a physical dexterity exam administered prior to employment (dex_i) and the years of education (lex_i).

$$y_i = \beta_0 + \beta_1 sex_i + \beta_2 dex_i + \beta_3 lex_i + \beta_4 lex_i^2 + \epsilon_i$$

With a total of 683 oberservations, the productivity level ranges from 10.5 to 19.1 and is unimodal and almost symmetric judging by the density plot. Our model investigates the "typical" (most likely) productivity levels given the covariates. We contrast our Bayesian MTD model with Yu and Aristodemu's semi-parametric Bayesian mode regression model that uses mixtures of symmetric uniform distributions. We use the first 675 observations as a training set to estimate the model. Two parallel chains with different initial values were run for the model as suggested by Yu and Aristodemu. The results are based on 500,000 iterations of which 300,000 samples were burned-in, and finally one in every 100 iterations was selected as the sample. Table 5 summarizes the parameter estimates and compares them with the ones from Yu and Aristodemu's model. While both models agree on the signs

[1]Full Year Consolidated Data File, Event Files and Medical Condition Files and Codebook, years 2002 to 2011. Agency for Healthcare Research and Quality. Center for Financing, Access and Cost Trends. Rockville, MD. November 2014.

Table 9: Summary Statistics of Model Parameters from MTD and Non-Parametric Mode Regressions (NBMR)

	MTD			NBMR		
Parameter	Mean	S. D.	95% CI	Mean	S. D.	95% CI
β_0	4.80	0.27	(4.25, 5.19)	4.10	1.11	(2.56, 6.52)
β_1	-0.80	0.11	(-1.06, -0.56)	-0.84	0.08	(-1.03, -0.73)
β_2	0.11	0.004	(0.10, 0.12)	0.12	0.005	(0.11, 0.12)
β_3	0.90	0.02	(0.87, 0.93)	1.08	0.18	(0.69, 1.37)
β_4	-0.04	0.002	(-0.042, -0.035)	-0.05	0.008	(-0.06, -0.03)

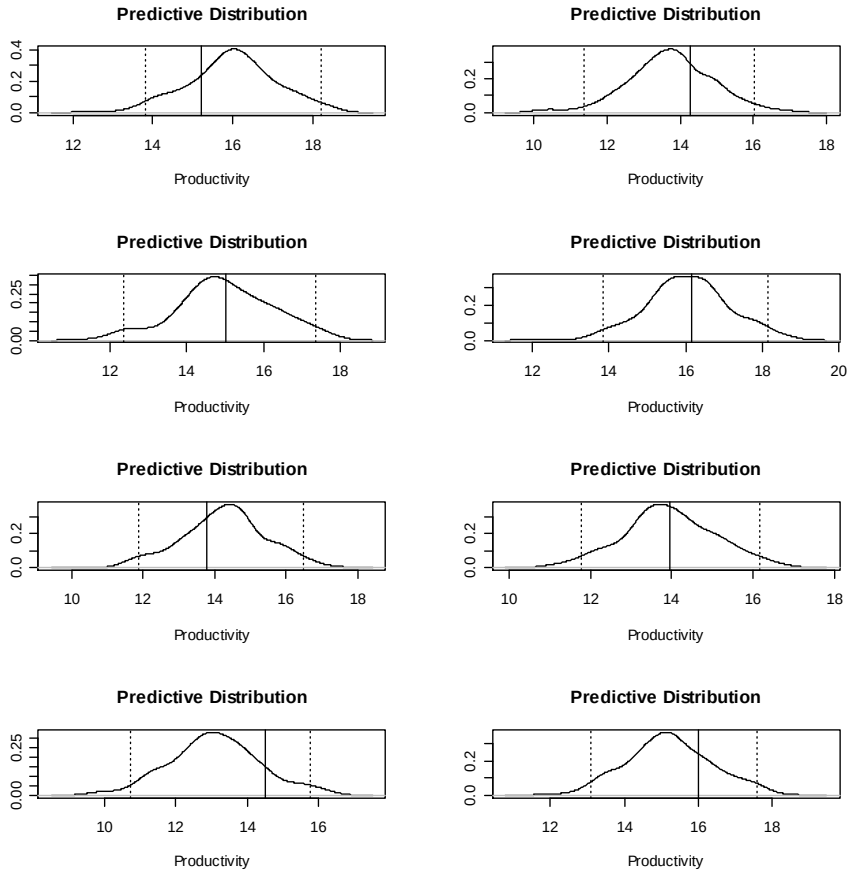
of the regression coefficients, our model provides a smaller credible interval for most of the parameters.

We use the last eight observations to validate the out-of-sample accuracy of our model. Figure 1 summarizes the predictive density plots for each of these last eight values, the true values, and the 95% predictive intervals. All the eight true values lie in their corresponding credible intervals. Also, Table 6 provides the numerical values corresponding to the point and interval predictions for these eight out-of-sample values.

Table 10: The Summary for the Test Sets

y	Predictive Mean	95% CI		sex	dex	lex	lex2
15.22000	15.980	13.830	18.198	0	52	12	144
14.27919	13.736	11.363	16.012	1	44	15.5	240.25
15.01482	14.998	12.355	17.364	0	43	12	144
16.17027	16.055	13.856	18.161	0	53	12	144
13.78203	14.251	11.852	16.493	0	38	12	144
13.94473	13.938	11.763	16.164	0	35	10	100
14.50122	13.106	10.729	15.771	1	37	14	196
16.00031	15.221	13.080	17.583	0	46	12.5	156.25

Figure 8: Preditive Distributions for Six Observations: The Solid Vertical Line is the Actualy Productivity and the Dotted Vertical Lines are the 95% Predictive Intervals.

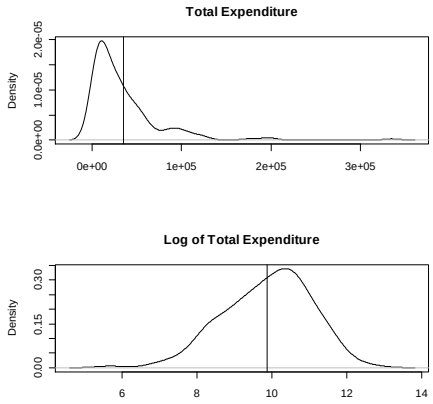


6.2 End Stage Renal Disease (ESRD) Model

This was a cross-sectional study of ESRD adults using a nationally representative Medical Expenditure Panel Survey (MEPS) from 2002 to 2011. MEPS is co-sponsored by the Agency of Healthcare Research and Quality and the National Center for Health Statistics (NCHS). The MEPS survey, initiated in 1996, has collected data annually that can be utilized to provide nationally representative estimates of the intensity, frequency, and the cost of healthcare services that Americans use and how these services are covered and paid for by different insurance providers. MEPS data can be accessed through the website administered by AHRQ at www.meps.ahrq.gov.

Zanwar (2012) used a portion of the MEPS data, which is what we too use in this illustration. The total sample size after carefully eliminating a multitude of recording errors resulted in 191 observations.

Figure 9: The Density Plot for TE andln(TE): The Solid Vertical Line Indicates the Mean



The response variable of interest in this study is the Total Expenditure (TE) incurred by each ESRD patient each year, which is the sum of the following expenses: ER, Inpatient, Outpatient, Office visits, Medical equipment/supply, Prescription drug, and Other home health care. In the time frame considered, it should be noted that almost half of the patients had only one record of TE while the rest had two. While a repeated measures GLM could be used to handle such data, for the purposes of illustrating our mode regression, we created a nominal variable, REP, that equals one for patients with two TE observations, zero otherwise. It should be noted that the database did not contain any information as to when the patients died. The key point of this example is to illustrate the MTD mode regression model for highly skewed data. MEPS also includes information on household income, medical conditions and clinical classification codes. Also, data on demographic and socioeconomic variables, such as gender, age, race, family income, region, insurance coverage, are available for respondents and their families residing in the U.S. We use four groups of covariates: demographic; disease type; types of services; and co-morbid conditions. Variable definitions are provided in Table 5.

TE in this dataset ranges from \$294 to \$335k. As shown in Figure 9, TE is highly right skewed with skewness 3.23 and after a natural log transformation, it becomes left skewed with skewness -0.42. We apply our MTD mode regression to analyze TE, and contrast it with a standard OLS-based multiple mean regression model. One could use a Bayesian mean regression model for comparison, but under a non-informative prior, the OLS and Bayes estimates are roughly the same.

- Based on subject matter knowledge and some preliminary data analysis such as graphical and correlation analyses, we narrowed the number of variables down to those given in the model

Table 11: The Code and Description of Variables in MEPS

#	Categories / Code	Lable
1	TE	Total Medical Related Expenditures
2	REP	1 if the patient has repeated measures
	Disease Type:	
3	TRANSP	Organ or Tissue Replaced by Transplant
4	DIALYSIS	Encounter for Dialysis and Dialysis Catheter Care
5	CKD	Chronical Kidney Disease
	(reference group)	Other Postprocedural States
	Region:	
6	NORTHE	if Patients from Northeast Region
7	MIDWEST	if Patients from Midwest Region
8	SOUTH	if Patients from South Region
	(reference group)	Patients from West Region
	Type of Services:	
9	HHNUM	# Home Health events association with condition
10	IPNUM	# of Inpatient Events Associated with Condition
11	OPNUM	# of Outpatient Events Associated with Condition
12	OBNUM	# of Office Based Events Associated with Condition
13	ERNUM	# of ER events associated with condition
14	RXNUM	# of Prescribed Medicines Associated with Condition
	Co-Morbid Conditions:	
15	DIAB	Diabetes Diagnosis, Round 5/3
16	HBP	High Blood Pressure Diagnosis ($>$ 17 Years), Round 5/3
17	CHD	Coronary Heart Disease Diagnosis ($>$ 17 years), Round 5/3
18	ANG	Angina Diagnosis ($>$ 17 years), Round 5/3
19	OTHR	Other Heart Disease Diagnosis ($>$ 17)
20	SMOKEIN	Current Smoker (SAQ weight)

Table 12: Regression Coefficients Summarization

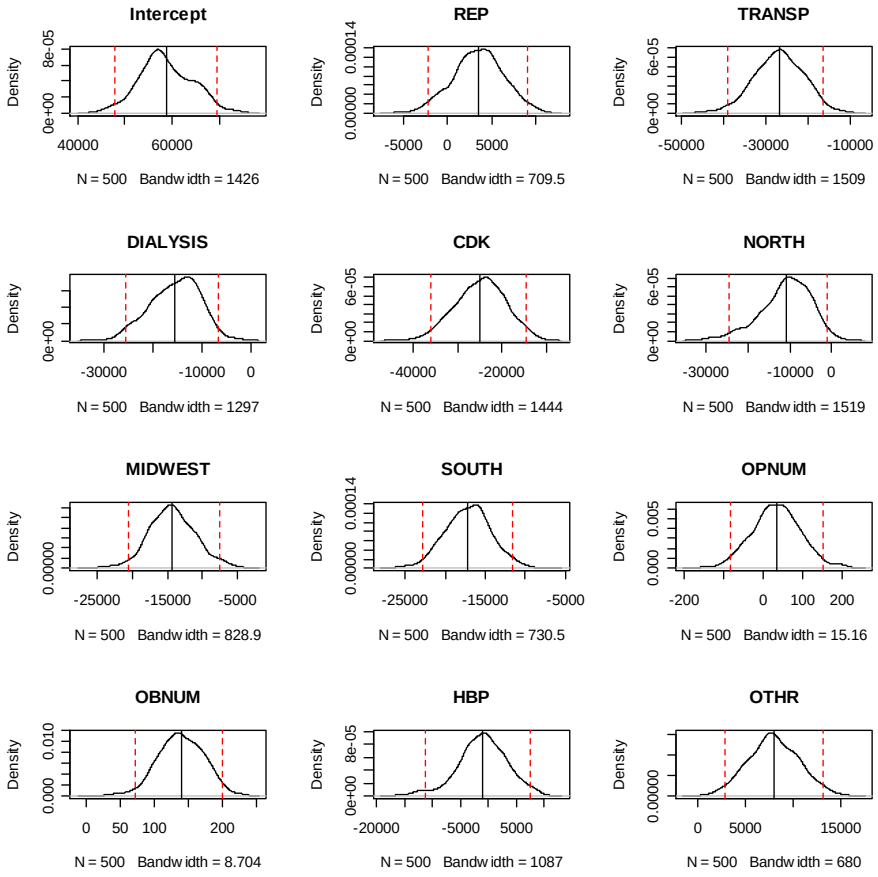
	OLS Mean Regression		MTD Mode Regression		
	Estimate	Std.Error	Posterior Mean	Upper CI	Lower CI
Intercept	17437.83	14356.23	58629.46	47974.52	69269.20
REP	-3546.87	5706.62	3500.64	-2109.86	9054.98
TRANSP	-22804.24	15469.98	-26823.11	-39031.77	-16195.52
DIALYSIS	-910.72	14569.21	-15439.68	-25444.81	-6777.71
CKD	-13549.30	14826.50	-24846.98	-35880.87	-14686.19
NORTHE	1118.12	10608.03	-10951.80	-24531.73	-1233.59
MIDWEST	-1058.38	7870.46	-14272.83	-20651.15	-7439.66
SOUTH	-9416.44	6755.52	-17130.47	-22892.86	-11515.00
OPNUM	360.32	111.45	33.91	-81.75	152.15
OBNUM	293.31	60.39	138.75	71.25	200.17
HBP	22679.60	9795.05	-1013.18	-11412.36	7420.72
OTHR	5894.23	5813.67	7996.54	2798.48	13139.29

below.

$$TE = \alpha + \beta_1 REP + \beta_2 TRANSP + \beta_3 DIALYSIS + \beta_4 CKD + \beta_5 NORTHE$$
$$+\beta_6 MIDWEST + \beta_7 SOUTH + \beta_8 OPNUM + \beta_9 OBNUM + \beta_{10} HBP + \beta_{11} OTHR$$

Consider Table 12. Compared to the West region, the other three regions have substantially lower costs. Also compared to other Postprocedural States, Transplant, Dialysis, and CKD have much lower costs. Each additional Office-based Visit adds roughly \$139 to the total cost and patients with Other Heart Disease paid roughly \$8000 more than patients without it. The posterior distributions for all the coefficients appears in Figure 10. For comparison purposes, Table 12 also includes the standard OLS estimates from a mean regression. Even after taking the natural logarithm of TE, the data are still skewed. Hence, the conditional mode is a better estimator of the central tendency for these highly skewed, unimodal data; see also, Collomb et al. (1987), Quintela-Del-Rio and Vieu (1997), Ould-Sad (1997), Berlinet et al. (1998), or Louani and Ould-sad (1999). This is because it is less influenced by outliers. And to be sure, there are quite a few outliers in the current dataset.

Figure 10: Posterior Distributions of the Regression Parameters: Red Dotted Lines Indicate the 95% Credible Intervals and the Black Solid Lines Indicate the Posterior Means



7 Bayesian Mode Univariate Dynamic Linear Models

Consider the following time series setup. At any time t, the observed data y_t is perturbed by noise (or error), u_t. The expected value of y_t is modeled via the regression term $X\beta$. The thrust of a Dynamic Linear Model (DLM) is to allow the β vector to evolve over time, typically as an autoregressive process with additional noise (or error) over time, say v_t; that is, β is now essentially β_t. The error u_t is called the observation error, while v_t is called the system error, leading to two inter-related equations, the Observation and System (or State) equations. It is convention to denote the regressors in the observation equation as F and the collection of system parameters as G; see West and Harrison (1997) and Petris et al. (2009). The observation equation is given by

$$y_t = F_t'\Theta_t + u_t \; u_t \sim N(0, \sigma_u^2), \; t = 1, ..., T$$

and the system equation is written as

$$\Theta_t = G_t\Theta_{t-k} + v_t \; v_t \sim N_q(0, V_t),$$

where u_t and v_t are assumed mutually independent and let Φ denote the collection of parameters corresponding to the regression functionsl and the error distributions.

This thesis develops the Bayesian updating and filtering equations for the above model when the distribution of u_t at each time period is triangular or MTD. With this assumption, the mean DLM changes to a mode DLM. In the following sections, a class of algorithms for non-Gaussian errors in state space models (SSM) namely sequential Monte Carlo (SMC) Methods will be introduced. Then, full Bayesian inference is made possible using particle Markov chain Monte Carlo (PMCMC) methods. This new class of models will be exemplified using simulated data.

7.1 Sequential Monte Carlo and Particle Markov Chain Monte Carlo Method

First introduced by Gordon et. al. (1993), the principle of Sequential Monte Carlo method is to approximate sequentially the posterior densities of states and marginal likelihoods of data given parameters. The sequence starts at the posterior distribution of the first state given the first obersvation $p_\Phi(\Theta_1|y_1)$ and the marginal likelihood of the first observation$p_\Phi(y_1)$, then $p_\Phi(\Theta_{1:2}|y_{1:2})$ and $p_\Phi(y_{1:2})$ and so on. These densities are approximated by a set of N weighted random samples called particles.

These particles evolve through the system equations and are adjusted by so-called importance weights associated with each particle. The advantage of SMC is that it is not restricted by assumptions of linearity or Gaussian noise. The principal drawbacks of SMC are that the efficiency of the approach depends on the number of particles, and the parameters are assumed to be known.

There are many different versions of SMC. In general, SMC is a combination of sequential importance sampling and resampling. The resampling part replaces the particles with relatively low weights with ones with higher weights. However, resampling introduces the problem of degeneracy. That is, the number of the unique value particles decreases as the number of the states to be estimated increases. Modifications such as Resample-Move method (Gilks and Berzuini, 2001) which essentially creates "jitter" through Markov kernels to reintroduce diversity, and Block Sampling (Doucet et. al. 2006) which only sample locally around a small number of states, are approaches to deal with the problem of degeneracy.

Another drawback of SMC is that the inference of parameters can only be made through the likelihood function. Andrieu et al (2010) proposed an algorithm, particle Markov Chain Monte Carlo method (PMCMC), that can address this issue by building efficient high dimensional proposal distributions through SMC and embed it within a standard MCMC algorithm. The main idea is to utilize SMC to create approximation densities for posterior densities of states variables, marginal densities of parameters and a Gibbs sampler for the joint posterior distribution of parameters and states. The advantage of this algorithm is that it makes the Bayesian inference for parameters feasible for a large class of non-Gaussian, non-linear state space models. In this thesis, our main focus is on the inference of the regression coefficients and the estimation of error distributions. We rely mainly on PMCMC to do Bayesian inference. It is possible there may be other competitive algorithms, but that's not the primary aim of this section; rather, the aim is to introduce a new class of mode DLMs using the triangular distribution for the error structure in the observation equation.

7.2 The Implementation of SMC

For the regression parameters in the state space model, canonically, consider the following:

$$\beta_1 \sim \mu_\theta(\cdot)$$

$$\beta_{n+1}|(\beta_n = b) \sim f_\theta(\cdot|b)$$

$$Y_n|(\beta_1, ..., \beta_n = b, ..., X_m) \sim g_\theta(\cdot|b) \qquad \text{for } 1 < n < m$$

At time 1, let $q_\theta(\beta_1|y_1)$ be the proposal density to approximate $p_\theta(\beta_1|y_1)$. We first sample N particles from $q_\theta(\beta_1|y_1)$ and compute and normalize the weights as follows,

$$w_1(\beta_1^k) := \frac{p_\theta(\beta_1^k, y_1)}{q_\theta(\beta_1^k, y_1)} = \frac{\mu_\theta(\beta_1^k) g_\theta(y_1|\beta_1^k)}{q_\theta(\beta_1^k, y_1)}, \tag{7.1}$$

$$W_1^k := \frac{w_1(\beta_1^k)}{\sum_{m=1}^N w_1(\beta_1^m)}. \tag{7.2}$$

Let $\mathbf{W_n} = (W_n^1, ..., W_n^N)$, denote the normalized weights at time n and $\mathcal{F}(\cdot|\mathbf{p})$ denotes the multinomial probability distribution on $\{1, ..., m\}$ with $\mathbf{p} = (p_{1,...}p_m)$ where $p_k \geq 0$ and $\sum p_k = 1$.

At times $n = 2, ...T$, first sample the index for each particle $A_{n-1}^k \sim \mathcal{F}(\cdot|\mathbf{W_{n-1}})$, sample $\beta_n^k \sim q(\cdot|y_n, \beta_{n-1}^{A_k^{n-1}})$, and set $\beta_{1:n}^k = (\beta_{1:n-1}^{A_k^{n-1}}, \beta_n^k)$. Then, compute and normalize the weights as follows,

$$w_n(\beta_{1:n}^k) := \frac{p_\theta(\beta_{1:n}^k, y_{1:n})}{p_\theta(\beta_{1:n-1}^k, y_{1:n-1}) q_\theta(\beta_n^k|y_n, \beta_{1:n-1}^{A_{n-1}^k})} = \frac{f_\theta(\beta_n^k|\beta_{1:n-1}^{A_{n-1}^k}) g_\theta(y_n|\beta_n^k)}{q_\theta(\beta_n^k|y_n, \beta_{1:n-1}^{A_{n-1}^k})}, \tag{7.3}$$

$$W_n^k := \frac{w_n(\beta_{1:n}^k)}{\sum_{m=1}^N w_n(\beta_{1:n}^m)}. \tag{7.4}$$

Let $\mathbf{A}_n := (A_n^1, ..., A_n^N)$, for any $n = 1, ..., T-1$. $\mathbf{A}_n$ is distributed as follows.

$$r(\mathbf{A}_{n-1}|\mathbf{W}_{n-1}) := \prod_{k=1}^N \mathcal{F}(A_{n-1}^k|\mathbf{W_{n-1}}).$$

One can intepret $\mathbf{A}_{n-1}$ as the aggregate index of which offspring particles at time n choose their ancestor particles at time $n-1$. Introduce a index variable B_n^k to denote the ancestor particle of $\beta_{1:T}^k$ at generation n. By fixing the last generation index $B_T^k := k$, for $n = 1, ..., T-1$, one can observe the backward recursive relation $B_n^k := A_n^{B_{n+1}^k}$. As a result for any $n = T-1, ..., 1$, we have the identity $\beta_{1:T}^k = (\beta_1^{B_1^k}, \beta_2^{B_2^k}, \ldots, \beta_{T-1}^{B_{T-1}^k}, \beta_T^{B_T^k})$ and $B_{1:T}^k = (B_1^k, B_2^k, ..., B_{T-1}^k, B_T^k = k)$ is the ancestral 'lineage' of a particle.

The SMC procedure provides an approixmation to the joint posterior density $p_\theta(\beta_{1:T}^k|y_{1:T})$ given

by

$$\hat{p}_\theta(\beta_{1:T}^k|y_{1:T}) := \sum_{k=1}^{N} W_T^k \delta_{\beta_{1:T}^k}(\mathrm{d}\beta_{1:T}).$$

One can draw samples from $p_\theta(\beta_{1:T}^k|y_{1:T})$ by drawing an index from the discrete distributions $\mathcal{F}(\cdot|\mathbf{W}_T)$. In addition, one can also approximate the marginal likelihood $p_\theta(y_{1:T})$ by

$$\hat{p}_\theta(y_{1:T}) := \hat{p}_\theta(y_1) \prod_{n=1}^{T} \hat{p}_\theta(y_n|y_{1:n-1})$$

where

$$\hat{p}_\theta(y_n|y_{1:n-1}) = \frac{1}{N} \sum_{k=1}^{N} w_n(\beta_{1:n}^k)$$

is an estimate, at time n, of

$$p_\theta(y_n|y_{1:n-1}) = \int w_n(\beta_{1:n}) q_\theta(\beta_n|y_n, \beta_{n-1}) p_\theta(\beta_{1:n-1}, y_{1:n-1}) \mathrm{d}\beta_{1:n}.$$

Note that the dependency of $w_n(\beta_{1:n})$ on $\beta_{1:n}$ is only through $\beta_{n-1:n}$. That is, the weights of the particles only depends on the current and previous generations of particles.

7.3 Gibbs Sampler for DLM with Triangular Errors

In this subsection, we develop the Gibbs sampler for a mode DLM assuming only one regressor in the observation equation and take the state equation to be a random walk process.

$$g_\theta(y_n|\beta_n) = \alpha + \beta_n x_n + \varepsilon_n \tag{7.5}$$

$$f_\theta(\beta_n|\beta_{n-1}) = \beta_{n-1} + \nu_n \; n = 2, ...T, \tag{7.6}$$

where ε_n follows a triangle distribution with parameters (a, λ), $\nu_n \sim N(0, \sigma^2)$ and $D_n = \{y_{1:n}, x_{1:n}\}$ are the observed data at time n. Also assume the first state follows a standard normal distribution $\beta_1 \sim N(0, 1)$. Let θ denote the model parameters including $(\alpha, a, \lambda, \sigma^2)$; then the posterior density of the states variables is given by

$$p_\theta(\beta_{1:T}|D_{1:T}) \propto p_\theta(\beta_{1:T},\, y_{1:T}|x_{1:T}) = \mu_\theta(\beta_1)\prod_{n=2}^{T} f_\theta(\beta_n|\beta_{n-1})\prod_{n=1}^{T} g_\theta(y_n|\beta_n)$$

and the joint posterior density of both states and parameters is

$$p(\theta, \beta_{1:T}|y_{1:T}) \propto p_\theta(\beta_{1:T},\, y_{1:T})p(\theta).$$

Assume independent priors for $p(\theta) = \pi_\alpha(\cdot)\pi_a(\cdot)\cdot\pi_\lambda(\cdot)\cdot\pi_{\sigma^2}(\cdot)$, standard normal for α, improper prior for λ, Pareto$(a\alpha_1, a\alpha_2)$ for a, and Inverse Gamma(IG)$(\sigma\alpha, \sigma\beta)$ for σ^2.

In short, the random quantities (states variables and parameters) for the entire DLM model that need to be estimated are $\{\alpha, \beta_{1:T}, a, \lambda, \sigma^2)$. Our goal is to construct an MCMC chain to sample these quantities sequentially from their corresponding conditional distributions. The particle Gibbs sampler proposed by Andrieu et al (2010) serves this situation best. The method consists of using a Gibbs sampler using iterative draws from $p(\theta|\beta_{1:T},\, y_{1:T})$ and $p_\theta(\beta_{1:T}|y_{1:T})$. The particle approximation to the Gibbs sampler requires the use of a special type of PMCMC update called the conditional SMC update. This update requires a prespecified path $\beta_{1:T}$ with ancestral lineage $B_{1:T}$ guranteed to survive all the resampling steps.

To obtain a sample from $p_\theta(\beta_{1:T}|y_{1:T})$, we first determine the ancestral lineage. Without any preference, we drawB_n, $n = 1, ..., N$ from a discrete uniform distribution on 1 to N. Set $\beta^B_{1:T} = \{\beta^{B_n}_n \sim N(\beta^{B_{n-1}}_{n-1}, \sigma^2, n = 1, ...T\}$. Use the prior density of the states as an importance density (Gordon et. al. 1993) and set $q_\theta(\beta_1|y_1) = \mu_\theta(\cdot)$, which is standard normal, and $q_\theta(\beta_n|y_n, \beta_{n-1}) = f_\theta(\beta_n|\beta_{n-1})$, which is normal with mean β_{n-1} and variance σ^2. The SMC procedure is very similar to the one in the previous subsection except that in each time n, we leave the particle that is specified by the ancestral lineage $\beta^{B_n}_n$. Specifically,

- For $n = 1$, sample $\beta^k_1 \sim N(0, 1)$ for $k \neq B_1$, compute $w_1(\beta^k_1)$ and normalize the weights $W^k_1 \propto w_1(\beta^k_1)$ according to (1.1) and (1.2)

$$w_1(\beta^k_1) := \frac{p_\theta(\beta^k_1, y_1)}{q_\theta(\beta^k_1, y_1)} = \frac{\mu_\theta(\beta^k_1)g_\theta(y_1|\beta^k_1)}{\mu_\theta(\beta^k_1)} = g_\theta(y_1|\beta^k_1) \tag{7.7}$$

$$W^k_1 := \frac{w_1(\beta^k_1)}{\sum_{m=1}^{N} w_1(\beta^m_1)}. \tag{7.8}$$

- For $n \geq 2$, for $k \neq B_n$, sample $A_{n-1}^k \sim \mathcal{F}(\cdot|\mathbf{W}_{n-1})$; sample $\beta_n^k \sim N(\beta_{n-1}^{A_{n-1}^k}, \sigma^2)$ for $k \neq B_1$; compute $w_n(\beta_{1:n}^k)$ and normalize the weights $W_n^k \propto w_n(\beta_{1:n}^k)$ according to (1.3) and (1.4)

$$w_n(\beta_{1:n}^k) := \frac{p_\theta(\beta_{1:n}^k, y_{1:n})}{p_\theta(\beta_{1:n-1}^k, y_{1:n-1}) q_\theta(\beta_n^k | y_n, \beta_{1:n-1}^{A_{n-1}^k})} = \frac{f_\theta(\beta_n^k | \beta_{1:n-1}^{A_{n-1}^k}) g_\theta(y_n|\beta_n^k)}{f_\theta(\beta_n^k | \beta_{1:n-1}^{A_{n-1}^k})} = g_\theta(y_n|\beta_n^k), \tag{7.9}$$

$$W_n^k := \frac{w_n(\beta_{1:n}^k)}{\sum_{m=1}^N w_n(\beta_{1:n}^m)}. \tag{7.10}$$

The full Gibbs sampler for the triangular model DLM is given by,

1. Initialization, $j = 0$, set $B_{1:T}(0)$, $\beta_{1:T}(0)$, $\sigma^2(0)$, $\alpha(0)$ arbitrarily, calculate $\varepsilon_{1:T}(0)$ and $\nu_{1:T}(0)$, set $a(0) = 2 \cdot \max(\varepsilon_{1:T}, -\varepsilon_{1:T})$ and $\lambda(0) = 0$

2. For iteration $j \geq 1$, sample the parameters $\theta(j) = (\alpha(j), a(j), \lambda(j), \sigma^2(j))$ and the auxiliary variable $v_{1:T}(j)$ given $\beta_{1:T}(j-1)$ as follows.

$$f(v_i|\varepsilon_i, a, \lambda) \propto \mathbf{1}(\max\{(\varepsilon_i)e^{-\lambda}, -(\varepsilon_i)e^{\lambda}\},\, a) \tag{7.11}$$

$$f(a|\mathbf{v}, \varepsilon_{1:T}, \lambda) \propto \prod_i f(v_i|a) \propto \pi(a) \cdot \frac{1}{a^{2n}} \mathbf{1}(a > \max v_i) \sim \text{Pareto}(\max(a\alpha_1, \mathbf{v}), , a\alpha_2 + 2T) \tag{7.12}$$

$$f(\lambda|\mathbf{v}, \varepsilon_{1:T}) \propto \left(\frac{1}{(2\cosh(\lambda)}\right)^n \mathbf{1}\left(\max\left(\ln(\frac{(\varepsilon_i)^+}{v_i})\right) < \lambda < \min\left(\ln(\frac{-v_i}{(\varepsilon_i)^-})\right)\right) \tag{7.13}$$

$$f(\alpha|\varepsilon_{1:T}, \beta_{1:T}, x_{1:T}, a, \lambda) \propto \exp(\frac{\alpha^2}{2}) \cdot \prod \mathbf{1}\left((y_i - \beta_i x_i) - v_i\lambda < \alpha < (y_i - \beta_i x_i) + v_i\lambda^{-1}\right)$$

$$f(\sigma^2|\beta_{1:T}, \nu_{1:T}) \sim \text{IG}(\sigma\alpha + \frac{T-1}{2}, \sigma\beta + \frac{\sum_{n=2}^T (\nu_n)^2}{2}).$$

Finally, run an SMC algorithm conditional on $\beta_{1:T}(j-1)$ and $B_{1:T}(j-1)$ to sample $\beta_{1:T}(j) \sim \hat{p}_{\theta(i)}(\cdot|y_{1:T})$ which involves drawing an index from the discrete distribution $\mathcal{F}(\cdot|\mathbf{W}_T)$.

7.4 Gibbs Sampler for DLM with MTD Errors

Now, if we assume ε_n follows MTD, we need to modify the calculation of the weight $g_\theta(y_1|\beta_1^k)$ and the PMCMC by changing the sampling of (a, λ, v_i) into the block $(a_j, \lambda_j, \delta_j, v_i, d_i, u_i)$. The former is hard to obtain since we do not nessasary get all components in the MTD and thus it is hard to compute the weights. This will be investigated as future research.

8 Future Research

A primary goal in the immediate future is to develop efficient simulation algorithms for Bayesian mode, univariate DLM regressions. The PMCMC algorithm in the last chapter converges very slowly. In order to better estimate the states variables, it requires a large amount of particles. Furthermore, sampling λ slows down the algorithm considerably.

One limitation of infinite MTD mode regression is that it is well-nigh impossible to obtain the likelihood function in any reasonable form. It would be nice to have a friendly analytical expression for the likelihood. For instance, one could ask how does a mode regression compare to a mean regression? One answer is the use of the Deviance Information Criterion. But this requires knowledge of the likelihood function.

One way to circumvent the likelihood issue is to use a finite mixture of MTD. In fact, it is easy to show that the excess kurtosis of a two-component MTD ranges from -0.6 to ∞. Some preliminary investigation for mode regression with a three-component MTD shows promise. Here a toy example illustrating the proposed methodology is given. Consider

$$y = \beta_1 + \beta_2 x + e$$

where $\beta_1 = 0$, $\beta_2 = 1$, x is drawn from a $\chi^2(3)$ distribution which is noramlized to have variance equaling 1, and e is distributed as a three-component MTD with mean 0.78 and standard deviation 1.7. The OLS estimator for (β_1, β_2) is (0.84,0.95) while the posterior mean from our mode regression is (-0.004, 1.005). Skewed data arises frequently in many applications, particularly finance. As we saw in the ESRD data analysis, even a log transformation may not resolve the problem with outliers. Many authors have therefore argued for the use of mode regressions.

One interesting future topic is to research the ideal number of MTD components to achieve both flexibilty and computational efficiency. It is also important to know whether the parameters in each component is identifiable. Unlike finite Gaussian mixtures where the components can be identified and traced by sorting the mean of each component, identifying the components of a finite MTD is difficult. One idea is to sort the components by a; however, the posterior mean of the parameters may not be consistent in this setup.

There is no existing research on variable selection for mode regression, and this is another topic for future research. It may be possible to modify the Stochastic Search Variable Selection (SSVS)

approach for mode regression.

Consider the linear mean regression model

$$Y = \sum_{j=1}^{m} \beta_j X_j + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

The setting for SSVS is as follows:

$$\beta_j|\gamma_j \sim (1-\gamma_j)N(0,\tau_j^2) + \gamma_j N(0, c_j^2\tau_j^2)\ c_j \gg 1$$

where $\gamma_j = 1$ indicates β_j has larger variance and hence has a higher chance to be different from 0, suggesting X_j should be included in the model. Assume independent priors for γ , where $f(\gamma) = \prod p_j^{\gamma_j}(1-p_j^{1-\gamma_j})$, then the posteior of γ is

$$P(\gamma_j = 1|\beta, \sigma, \gamma_{-j}) = \frac{c_1}{c_1 + c_2}$$

where

$$c_1 = f(\beta|\gamma_j = 1, \gamma_{-j}) \times f(\sigma|\gamma_j = 1, \gamma_{-j}) \times p_j$$

$$c_2 = f(\beta|\gamma_j = 0, \gamma_{-j}) \times f(\sigma|\gamma_j = 0, \gamma_{-j}) \times (1-p_j)$$

If the regression error term is assumed to follow a triangular distribution with parameter (a, λ) then,

$$c_1 = f(\beta|\gamma_j = 1, \gamma_{-j}) \times f(a, \lambda|\gamma_j = 1, \gamma_{-j}) \times p_j$$

$$c_2 = f(\beta|\gamma_j = 0, \gamma_{-j}) \times f(a, \lambda|\gamma_j = 0, \gamma_{-j}) \times (1-p_j)$$

Given γ_j the sampling for (β, a, λ) was studied in this thesis. To implement the modified SSVS one has to sample $c_1$and c_2 which involves the likelihood for the triangular distribution.

Finally, this thesis only discusses the univariate, mode regression model. Extending the ideas to more than one dimension is a daunting challenge, since it would require developing the theory of triangular distributions in high dimensions.

References

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (2003), "Posterior consistency for semi-parametric regression problems." *Bernoulli*, 9, 291–312.

Bannerman-Thompson, H. (2008), *Convex Density Function Estimation*, ProQuest.

Barker, W., Cuypers, M., and Holden, L. (2001), "Quantifying uncertainty in production forecasts: another look at the PUNQ-S3 problem." *Society of Petroleum Engineers Journal*, 6, 433–441.

Berlinet, A., Vajda, I., and der Meulen, E. V. (1998), "About the asymptotic accuracy of Barron density estimates." *Information Theory, IEEE Transactions*, 44, 999–1009.

Besag, J. and Green, P. J. (1993), "Spatial statistics and Bayesian computation." *Journal of the Royal Statistical Society. Series B (Methodological)*, 55, 25–37.

Bickel, P. and Fan, J. (1996), "Some problems on the estimation of unimodal densities." *Statistica Sinica*, 6, 23–46.

Birg'e, L. (1997), "Estimation of unimodal densities without smoothness assumptions." *The Annals of Statistics*, 25, 970–981.

Brunner, L. (1992), "Bayesian nonparametric methods for data from a unimodal density." *Statistics & Probability letters*, 14, 195–199.

Cai, B., Meyer, R., and Perron, F. (2008), "Metropolis–Hastings algorithms with adaptive proposals," *Statistics and Computing*, 18, 421–433.

Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992), "A Monte Carlo approach to nonnormal and nonlinear state-space modeling." *Journal of the American Statistical Association*, 87, 493–500.

Chernoff, H. (1964), "Estimation of the mode." *Annals of the Institute of Statistical Mathematics*, 16, 31–41.

Collomb, G., Hardle, W., and Hassani, S. (1987), "A note on prediction via estimation of the conditional mode function." *Journal of Statistical Planning and Inference*, 15, 227–236.

Damien, P., Wakefield, J., and Walker, S. (1999), "Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61, 331–344.

Dunson, D., Pillai, N., and Park, J. (2007), "Bayesian density regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 163–183.

Eddy, W. (1980), "Optimum kernel estimators of the mode," *The Annals of Statistics*, 8, 870–882.

Ferguson, T. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1, 209–230.

Floris, F., Bush, M., Cuypers, M., Roggero, F., and Syversveen, A.-R. (2001), "Methods for quantifying the uncertainty of production forecasts: a comparative study." *Petroleum Geoscience*, 7, S87–S96.

Gasser, T., Hall, P., and Presnell, B. (1998), "Nonparametric estimation of the mode of a distribution of random curves." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 681–691.

Gelfand, A. and Smith, A. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American statistical association*, 85, 398–409.

Gilks, W. and Wild, P. (1992), "Adaptive rejection sampling for Gibbs sampling," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41, 337–348.

Grenander, U. (1965), "Some direct estimates of the mode." *The Annals of Mathematical Statistics*, 36, 131–138.

Hall, P. and Huang, L. (2001), "Nonparametric kernel regression subject to monotonicity constraints." *The Annals of Statistics*, 29, 624–647.

Hall, P., Melville, G., and Walsh, A. H. (2001), "Bias correction and bootstrap methods for a spatial sampling scheme," *Bernoulli*, 7, 829–846.

Heckman, D., Geiser, D., Eidell, B., Stauffer, R., Kardos, N., and Hedges, S. (2001), "Molecular evidence for the early colonization of land by fungi and plants." *Science*, 293, 1129.

Hedges, S. and Shah, P. (2003), "Comparison of mode estimation methods and application in molecular clock analysis." *BMC Bioinformatics*, 4, 31.

Higdon, D. (1998), "Auxiliary variable methods for Markov chain Monte Carlo with applications," *Journal of the American Statistical Association*, 93, 585–589.

Ho, M.-W. (2006), "Bayesian density regression." *Journal of Computational and Graphical Statistics*, 15, 848–860.

Hong, Y., Tu, J., and Zhou, G. (2007), "Asymmetry in stock returns: statistical tests and economical evaluation." *Review of Financial Studies*, 20, 1547–1581.

Kalli, M., Griffin, J., and Walker, S. (2011), "Slice sampling mixture models," *Statistics and Computing*, 21, 93–105.

Kemp, G. and Silva, J. (2012), "Regression towards the mode." 170, 92–101, journal of Econometrics.

Kim, J. K. and Pollard, D. (1990), "Cube-Root asymptotics." *Annals of Statistics*, 18, 191–219.

Kumar, S. and Hedges, S. (1998), "A molecular timescale for vertebrate evolution." *Nature*, 392, 917–920.

Law, A. and Kelton, W. (2000), *Simulation Modeling and Analysis*, Boston: McGraw-Hill.

Lee, M. (1989), "Mode regression." *Journal of Econometrics*, 42, 337–349.

— (1993), "Quadratic mode regression." *Journal of Econometrics*, 57, 1–19.

Lo, A. (1984), "On a class of Bayesian nonparametric estimates: I. Density estimates," *The Annals of Statistics*, 12, 351–357.

Louani, D. and Ould-sad, E. (1999), "Asymptotic normality of kernel estimators of the conditional mode under strong mixing hypothesis," *Journal of Nonparametric Statistics*, 11, 413–442.

Markov, H., Valtchev, T., Borissova, J., and Golev, V. (1997), "An algorithm to "clean" close stellar companions." *Astronomy and Astrophysics Supplement Series*, 122, 193–199.

McVinish, R., Rousseau, J., and Mengersen, K. (2009), "Bayesian goodness of fit testing with mixtures of triangular distributions," *Biometrics*, 36, 337–354.

Meyer, M. (2001), "An alternative unimodal density estimator with a consistent estimate of the mode." *Statistica Sinica*, 11, 1159–1174.

Neal, R. M. (2003), "Slice sampling." *The Annals of Statistics*, 31, 705–741.

Ould-Sad, E. (1997), "A note on ergodic processes prediction via estimation of the conditional mode function," *Scandinavian journal of statistics*, 24, 231–239.

Parzen, E. (1962), "On estimation of a probability density function and mode." *The Annals of Mathematical Statistics*, 33, 1065–1076.

Patel, H., Sircar, A., Sheth, S., and Jadvani, R. (2011), "Application of genetic algorithm to hydrocarbon resource estimation," *Journal of Petroleum and Gas Engineering*, 2, 83–92.

Perron, F. and Mengersen, K. (2001), "Bayesian nonparametric modeling using mixtures of triangular distributions." *Biometrics*, 57, 518–528.

Petris, G., Petrone, S., and Campagnoli, P. (2009), *Dynamic Linear Models with R*, London: Springer.

Quintela-Del-Rio, A. and Vieu, P. (1997), "A nonparametric conditional mode estimate." *Journal of Nonparametric Statistics*, 8, 253–266.

Rao, C. (2010), "Report on application of probability in risk analysis in oil and gas Industry." .

Samanta, M. and Thavaneswarn, A. (1990), "Non-parametric estimation of the conditional mode." *Communications in Statistics Theory and Methods*, 19, 4515–4524.

Scherer, W. T., Pomroy, T. A., and Fuller, D. N. (2003), "The triangular density to approximate the normal density: decision rules-of-thumb." *Reliability Engineering & System Safety*, 82, 331–341.

Sethuraman, J. (1994), "A constructive definition of Dirichlet process priors," *Statistica Sinica*, 4, 639–650.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman & Hall.

Smith, M. and Kohn, R. (1997), "A Bayesian approach to nonparametric bivariate regression." *Journal of the American Statistical Association*, 92, 1522–1535.

van de Geer, S. (1993), "Hellinger-consistency of certain nonparametric maximum likelihood estimators." *The Annals of Statistics*, 14–44.

Walker, S. (2007), "Sampling the Dirichlet mixture model with slices," *Communications in Statistics: Simulation and Computation*, 36, 45–54.

Walker, S. and Hjort, N. L. (2001), "On Bayesian consistency." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 811–821.

West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models.*, New York: Springer-Verlag.

Wood, S. and Kohn, R. (1998), "A Bayesian approach to robust binary nonparametric regression." *Journal of the American Statistical Association*, 93, 203–213.

Yasukawa, K. (1926), "On the probable error of the mode of skew frequency distributions." *Biometrika*, 18, 263–292.

Yu, K. and Aristodemou, K. (2014), "Bayesian mode regression." Submitted for publication.

Zanwar, P. (2012), "Effects of overweight and obesity on economic outcomes among U.S. adults with kidney disease." Ph.D. thesis, Texas Medical Center Dissertations.

Zhang, G. (2003), "Estimating uncertainties in integrated reservoir studies." Ph.D. thesis, Texas A & M University.

Ziegler, K. (2003), "On the asymptotic normality of kernel regression estimators of the mode in the random design model." *Journal of Statistical Planning and Inference*, 115, 123–144.